

AD 714092

RESEARCH IN ON-LINE COMPUTATION

by

David O. Harris, James A. Howard, Roger C. Wood

University of California
Santa Barbara, California 93106

Contract No. AF19(628)-6004
Project No. 8684

FINAL REPORT

26 April 1966 - 30 June 1970

Date of report 30 June 1970

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U. S. Government.

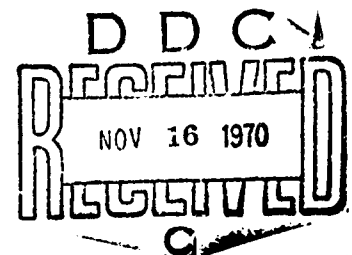
This document has been approved for public release and sale; its distribution is unlimited.

Details of Restrictions in
this document may be better
stated on microfiche

Contract Monitor: Hans H. Zschirnt
Data Sciences Laboratory

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
Springfield, Va 22151

Sponsored by
Advanced Research Projects Agency
ARPA Order No. 865
Monitored by
AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS 01730



Program Code No.9D30
Effective Date of Contract3 May 1966
Contract Expiration Date.....30 June 1970
Principal Investigator and Phone No.Dr. David O. Harris/ 805 961-2534
Project Scientist or Engineer and Phone No.Dr. Hans H. Zschirnt/ 617 861-3671

ACCESSION for	
CFSTI	WHITE SECTION <input checked="" type="checkbox"/>
SEC	PUFF SECTION <input type="checkbox"/>
UNCLASSIFIED	<input type="checkbox"/>
RESTRICTION	
BY	
DISTRIBUTION AVAILABILITY CODE	
DISC.	AVAIL. AND X CODE
/	

Qualified requestors may obtain additional copies from the Defense Documentation Center. All others should apply to the Clearinghouse for Federal Scientific and Technical Information.

ABSTRACT

The report covers the on-line computing system development from 1966 through 1970. It includes a general resume of progress through December, 1969 and a detailed progress from then through June 30, 1970. The improved version of the on-line system substantially improves system reliability and presents users new options. Significant progress in speech analysis/synthesis project includes: improved techniques for deriving accurate data from ASCON parameters, good results from the steady-state vowel recognizer, and one-pass analysis and synthesis. The 1800 has been improved so that it is a more effective research tool supporting the speech effort.

TABLE OF CONTENTS

PART I

Introduction.....	i
List of Scientists and Engineers.....	i
List of Publications and Reports.....	i
Related Research.....	iii
Resume.....	iv

PART II

Detailed Report of Period January through June 1970

Synopsis.....	1
Technical Findings and Major Accomplishments.....	2
A. Software.....	2
(1) ARPA Network.....	2
(2) UCSB On-Line System.....	3
B. 360/75 On-Line System, Hardware.....	4
C. Speech Project, General.....	7
Speech Project, Theory.....	7
(a) Wave-Function Structure of English Phonemes.....	8
(b) Preprocessing of Acoustic Waveform.....	13
(c) Improved Wave-Function Analysis/Synthesis System.....	23
(d) Computer Classification and Recognition of Phonetic Information.....	49
(e) Data Compression Studies.....	74
(f) Interrelationships between a Wave-Function Representation and a Formant Model of Speech.....	77
Speech Project, Software and Hardware.....	91
(a) On-Line System for Biological Research.....	92
(b) SEL-810B Software.....	97
(c) SEL-810B Interface to Speech Station.....	100
Conclusion.....	105

PART I

Introduction

This final report consists of a rather general resume of technical progress which has previously been reported in detail during the period April 26, 1966 through December 2, 1969. Following this resume is a detailed report covering the period from December 3, 1969 through June 30, 1970, the final reporting period for the current contract. From the report it will be clear that further research is indicated. Continuation of this research will be accomplished under Contract AF19628-70-C-0314 commencing July 1, 1970.

List of Scientists and Engineers Contributing to the Research

Dr. Glen J. Culler
Dr. David O. Harris
Dr. James A. Howard
Dr. Roger C. Wood
Mr. Roland F. Bryan
Mr. Ronald Stoughton

List of Publications and Reports Resulting from Sponsorship of the Contract

Publications

1. Baldwin, Jr., John A., and Glen J. Culler, "Wall-Pinning Model of Magnetic Hysteresis", Journal of Applied Physics, Vol. 40, No. 7, June 1969, pp. 2828 - 2835.

2. Bruch, Jr., John C. and Roger C. Wood, "The Teaching of Hydrodynamics Using Computer Generated Displays", Bull. Mech. Engng. Educ., Vol. 1, Pergamon Press 1962, pp. 1 - 11.
3. Culler, Glen J., "Mathematical Laboratories: A New Power for the Physical Sciences", Interactive Systems for Exp. Applied Mathematics, Academic Press Inc., New York, 1968, pp. 355- 384.
4. Culler, Glen J., "On the Polar Equations for Linear Systems and Related Nonlinear Matrix Differential Equations", Transactions of the American Mathematical Society, Vol. 118, Issue 6, June, 1965, pp. 390-405.
5. Davenport, Demorest, Glen J. Culler, Richard B. Forward, and William G. Hand, "The Investigation of the Behavior of Microorganisms by Computerized Television", IEEE Transactions on Bio-Medical Engineering, Vol. BME-17, No. 3, July, 1970, pp. 230-237.
6. Hendren, Philip, Experiments in Forms, Using Computer Graphics, Sept., 1968.
7. Howard, James A., and Keith L. Doty, UCSB On-Line System Manual, Feb., 1969.
8. Howard, James A., Roger C. Wood, "Hybrid Simulation of Speech Waveforms Utilizing a Gaussian Wave Function Representation", Simulation, Sept., 1968, pp. 117-124.
9. Wood, Roger C. and Philip Hendron, "A Flexible Computer Graphic System for Architectural Design", Information Display, March/April, 1968, pp. 35-40.

Reports

1. First Quarterly Report, Reporting Period: April 16, 1966 - July 15, 1966.
2. Second Quarterly Report, Reporting Period: July 16, 1966 - October 15, 1966.
3. Third Quarterly Report, Reporting Period: October 16, 1966 - Jan. 15, 1967.
4. Fourth Quarterly Report, Reporting Period: Jan 16, 1967 - April 15, 1967.
5. Fifth Quarterly Report, Reporting Period: April 16, 1967 - July 15, 1967.

6. Sixth Quarterly Report, Reporting Period: July 16 - October 15, 1967.
7. Seventh Quarterly Report, Reporting Period: Oct. 16, 1967 - January 15, 1968.
8. Eighth Quarterly Report, Reporting Period: Jan. 16, 1968 - April 15, 1968.
9. Ninth Quarterly Report, Reporting Period: April 16, 1968 - July 15, 1968.
10. Tenth Quarterly Report, Reporting Period: July 16, 1968 - Oct. 15, 1968.
11. Eleventh Quarterly Report, Reporting Period: Oct. 16, 1968 - Jan. 15, 1969.
12. Twelfth Quarterly Report, Reporting Period: Jan. 16, 1969 - April 15, 1969.
13. Thirteenth Quarterly Report, Reporting Period: April 16, 1969 - July 15, 1969.
14. Semiannual Technical Report, Reporting Period: June 2, 1969 - Dec. 2, 1969.

Related Research - List of Previous and Related Contracts

The development of on-line computation at the University of California at Santa Barbara was initiated with the delivery of a gift from the Bunker-Ramo Corporation consisting of one Teleputer Control unit, one Data Set Control unit, and one storage tube display device. This was used to carry out:

NONR 4222(09):

Pilot Experiment - Our pilot experimental program consisted of utilizing the Teleputer console which was donated by the Bunker-Ramo Corporation located in the Computer Center of the University of California at Santa Barbara but tied to a leased telephone line feeding into the RW 400, AN/SFQ 27 equipment at the Bunker-Ramo Corporation in Canoga Park. Through this program we demonstrated that an adequate curvilinear display was possible over a conventional 201B data set. We developed the basic software underlying our present on-line system.

ARPA SD 319:

An Experimental Communication Laboratory - We designed and constructed a 16 station computer classroom and the associated time-sharing software

which is now being used by the Electrical Engineering Department, the Mathematics Department, the Chemistry Department, and long line at Harvard Computation Laboratory, the University of California, Los Angeles Physics Department, the University of Kansas, and at the Livermore Radiation Laboratory in Livermore, California.

NSF GP 5382:

Mathematical Applications of On-Line Computation - We designed and constructed a logical interface connecting the IBM 1311 Model 1 disk drive to our on-line system. We adapted our on-line software to include external users. We initiated mathematical research in the areas of non-linear integral equations and complex function theory.

NSF GJ 115 and GJ 693:

Development of an on-line computer network for Chemistry Education - This network ties none other universities from across the nation into the UCSB On-Line Computer System. The first station was operational in March, 1970. To date results of this network have been extremely successful.

Resume of Technical Progress

Work under the contract commenced April 26, 1966. Detailed technical progress has previously been reported as indicated in the prior listing of reports.

The general purpose of the research was to develop on-line computing. Specific tasks were to develop a modern computing system, establish a campus network, enhance human-computer communications and establish a national network for appropriate institutions. The resume will discuss

progress in each of these task areas.

Software was written to effectively share/transfer control between the on-line system and standard batch processing. In addition sub-programs were developed for the vast number of macros required for effective use of the system. As a normal development progressed the operating system changed from DOS to OS. Various languages were added to the system to enhance user options and make the system easier from the user point of view. Special features include entering and manipulating jobs in the batch mode from remote terminals (RJE). To improve system reliability and make the system more exportable a new version of the system software was developed and installed during the final year of the contract.

Hardware developments have included developing the On-Line Computing System within the 360/50 which was replaced by the 360/65, which was replaced by the current 360/75. Network activity has grown from a small nucleus of campus terminals to a national network supported by an NSF grant and includes preliminary operations of the still-growing ARPA network. To support the networks the necessary interfaces, buffer and multiplexor have been developed. Within the Electrical Engineering Department a computer classroom has been established consisting of sixteen stations, and several smaller classrooms have also been installed elsewhere on campus. Peripheral hardware has included development of the double keyboard, a blackboard plotter, a system employing a Grafacon tablet, a high speed buffer, and a bugwatcher to facilitate use of the computer in support of bio-medical engineering research.

The objective of the speech project is to establish effective human communications with the computer. Early efforts have been devoted to identifying the various elements of human speech and analyzing those that

could be useful in this communication. Research was fruitful in that the concept of the waveform analysis and synthesis approach has been fully developed. The parameters have been described, techniques and procedures outlined and essential hardware obtained to test the fundamental elements of speech sounds and to provide clear reproduction of these sounds. As a natural outgrowth of the continuing research effort, early techniques, procedures and hardware will be modified to enhance the reliability, improve efficiency, and develop applications.

PART II - Detailed Technical Progress for the Period January - June, 1970.

Synopsis

Network software development has progressed in consonance with network protocol development.

A new version of the basic system software was virtually completed during this reporting period. Operational testing under the rigors of normal user activity will commence on July 1, 1970.

Hardware development includes the Multi-Teletype Control prototype and the High Speed Data Buffer. Both units adhere to the concept of connecting directly to the 360 without going through the UCSB Buffer unit. The Multi-Line Controller is in the design phase and should be developed and operational during the next reporting period.

Significant progress has been made in the speech analysis/synthesis project. This includes: (1) Development of a one-pass analysis/synthesis system which substantially increases data accuracy. (2) Improved techniques for deriving reliable recognition information from ASCON parameters. (3) Achievement of good results from the steady-state vowel recognizer. The entire speech project has been enhanced considerably by improvements to the 1800 which is now a more useful tool with the capability of handling a substantial portion of the computing required in the continuing research effort.

Technical Findings and Major Accomplishments

A. Software

(1) ARPA Network

As a result of network protocol modifications it was necessary to set aside previous work on the network software and virtually start over again on May 1, 1970. In planning the new implementation, it was decided to make it independent of the On-Line System, giving batch-mode tasks (as well as On-Line System users) direct access to the Network.

In the new implementation, tasks communicate with the Network Control Program (NCP) by means of a supervisor call. One supervisor call routine suffices to perform all necessary network functions; a branch index passed to the routine specifies the desired function. In general, the supervisor call routine returns control to the invoking task upon initiation of the operation. The operation is completed by the I/O interrupt handler, which posts the event complete associated with the requesting task. This method of signaling the completion of an event was chosen as a powerful alternative to "blocking" the task until completion (as proposed in the Network literature), and makes feasible the eventual use of the supervisor call by subroutines of the On-Line System.

Currently, the NCP runs as a normal task in batch mode under HASP and the Operating System. Upon start of execution, an initialization routine, by making the necessary modifications to low core, initiates the NCP as both the I/O and supervisor call first-level interrupt handlers (FLIH), permitting it to (1) process I/O interrupts from the IMP, and (2) gain control when the Network supervisor call is issued by any task in the machine. Should the

NCP abnormally terminate, low core is returned to its original state, and OS's FLIH's are reinstated. In this manner, the software can be developed and debugged with no modifications to the operating system. Once it has reached operational status, the NCP can either be made a resident part of OS, or run as an extension of the Logger using the present technique.

At present, those routines which transfer data between sockets (READ and WRITE) are operational, and transfers between processes in the 360/75 have been made using supervisor calls. With the adoption of an official Host-Host protocol scheduled for July 13, 1970, those routines which establish, switch, and break connections will be developed.

(2) UCS8 On-Line System

One significant change to the improved system software, was to include trailing predicates in the new version. This reverses the action intended when the last technical report was written. Rationale for including trailing predicates was primarily based upon the fact that numerous users were employing this feature and to eliminate it would create a great deal of anguish among these users. It was felt the slight additional software overhead would be much easier than re-educating the entire user population.

Development of the new version is virtually completed. Many elements have been satisfactorily checked by development personnel. These verifications foster optimism; however the acid test will come with the deluge of programs and operations of normal system use by our user group. The target date for normal operational use of the new system is July 1, 1970. Because of this fact, the software portion of this report is abbreviated - more extensive details will be included in the next report after the new system has been in operational use.

B. 360/75 On-Line System, Hardware

In the last technical report a new direction for system development was set forth. Development was to proceed away from further attachment of special devices to the existing UCSB Buffer and toward directly addressable I/O devices attached to the Multiplexor Channel of the System/360. The attachment of the IMP was done in this manner. Figure B-1 shows the present hardware distribution.*

Implementation of hardware for direct attachment is presently underway and none has been operated as yet. However, fundamental changes in the software have been made to allow the addition of the new devices by direct attachment, when the hardware is completed.

The last report discussed the use of the UCSB Buffer as a "test-bed" for new devices while maintaining its present operation. In this way new devices would be attached to the existing Buffer for test until the direct attachment facilities exist. Two such attachments are underway. The first is the Multi-Teletype controller that will operate Teletypes located around the campus, the second is the modification of an existing segment on the Buffer to allow half-duplex operation with an acoustic coupler unit. Both of these devices will subsequently be attached to future hardware for direct program operation by the 360.

The following items summarize the position of the several projects underway:

- (1) The Multi-Teletype Control prototype unit has been tested on the

* This Figure was included in the last technical report and is presented again for reference.

existing Buffer. Fabrication has been completed on the final unit and tests will proceed using the UCSB Buffer until direct attachment is achieved.

(2) The High-Speed bulk data buffer is presently in check-out. A special direct attachment will be implemented for use on the System/360.

(3) The direct attachment facility will be gained through use of a Multi-Line Controller which is presently in the design phase. When completed, the Multi-Teletype Control, additional display consoles, a program settable time interval controller, remote computers, and remote job entry units will be attached.

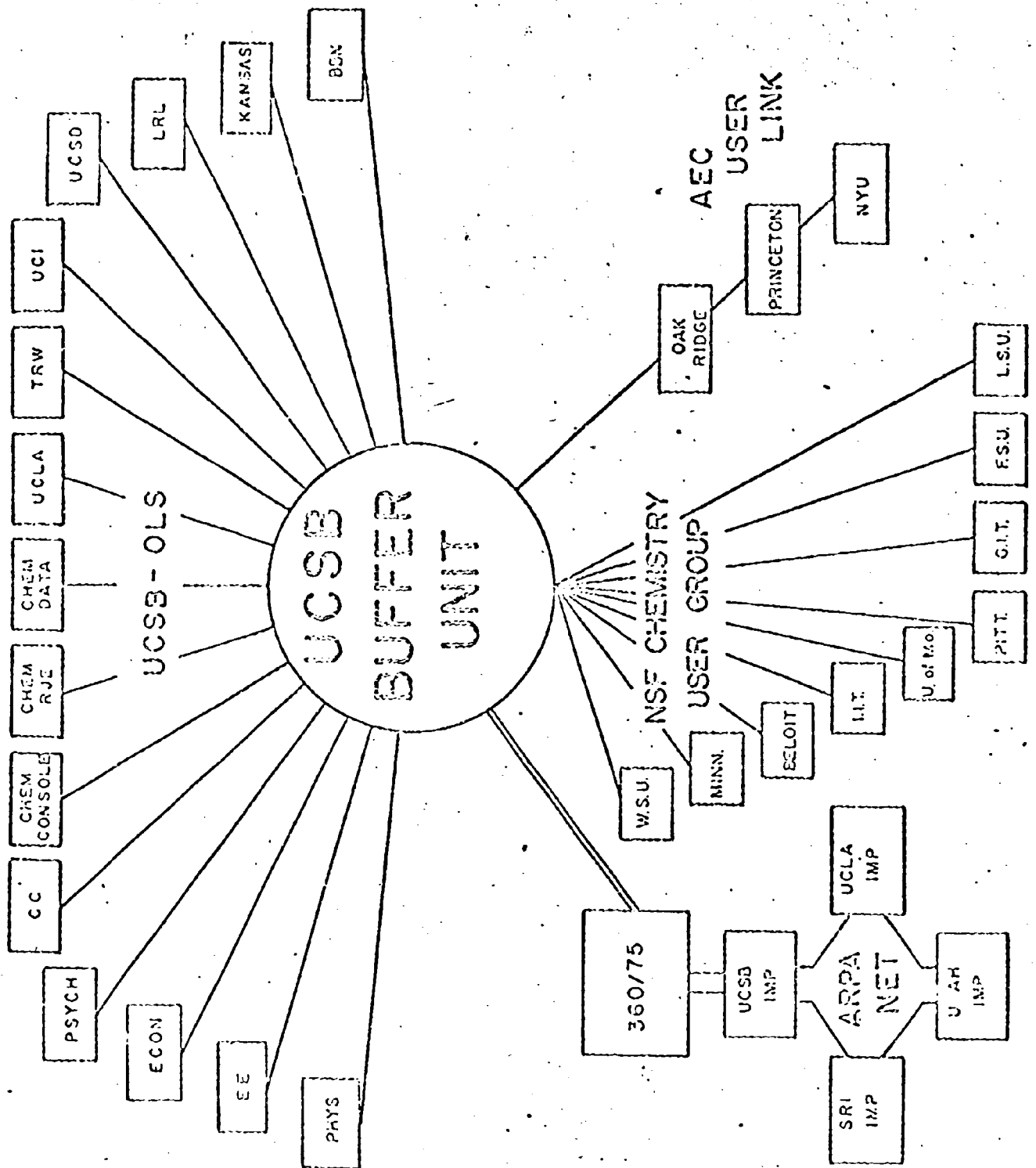


Figure B-1

C. Speech Project, General

The progress of the speech project is discussed in the following sections under the headings theory, software, and hardware respectively.

Speech Project, Theory

The theoretical aspects of the speech program were concerned with the following areas during this period:

- (1) Examination of the wave-function structure of the phonemes of the English language to define the requirements of the wave-function model of human speech.
- (2) Definition of a preprocessing method to filter the raw speech string into sub-strings amenable to wave-function analysis.
- (3) Development of a one-pass wave-function analysis/synthesis system based on the Gaussian Cosine Modulation (GCM) Model, that will accurately analyze and synthesize both male and female speech data.

The analysis programs have been structured to provide parameters that are compatible with the work being done on speech recognition (for example precision frequency information).

- (4) Continued studies on the computer classification and recognition of phonetic information including extraction of recognition parameters from the ASC ϕ N parameter set, recognition of steady-state vowels and vowels embedded between two unvoiced phonemes for a single speaker and preliminary studies of the segmentation of connected phonemes.

- (5) Studies of the data rate of the basic ASC ϕ N representation and the amount of data compression possible through the elimination of redundant wave-function sets.

(6) Preliminary definition of the interrelationship between the wave-function representation and a classical formant model of human speech.

Empirical formulae have been developed relating the ASCON parameters to formant amplitudes, frequencies, and bandwidths.

The above topics are discussed in detail in the following sections.

(a) Wave-Function Structure of English Phonemes

The success of an analysis system based on the Gaussian wave-function representation depends upon the accuracy with which the model covers the set of wave-functions found in filtered human speech. To verify the completeness of the model the wave-function structure of each of the 34 basic English phonemes, for a male and female voice, was studied. Two different filtering methods were employed for preprocessing the raw acoustic data into sub-strings amenable to wave-function analysis. These were as follows:

1. Adaptive Filtering - Filtering the raw speech data around the formants of the short-term energy spectrum as described in the previous semi-annual report.

2. Fixed Filtering - Filtering the speech data into four contiguous frequency bands covering the frequency range 100 - 3600 Hz in which the filter characteristics are fixed (identical) for all phonemes. Each of these filtering methods gave equivalent results. They both showed that the wave function model is not complete in the sense that it covers all wave functions found in filtered human speech. There are actually two separate classes of wave-functions of which the acoustic waveform may be composed.

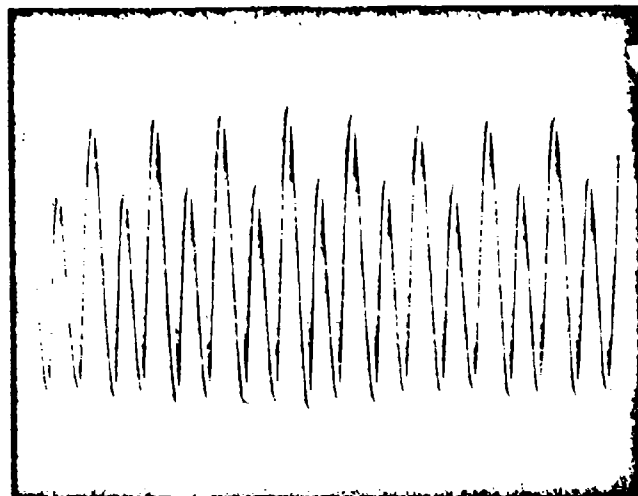
1. Waveforms with a Gaussian Envelope - The cyclic behavior of the waveform under the envelope may be described by either some appropriate Hermite

polynomial (Gaussian Wave-Function Model) or an appropriate cosine function (Gaussian Cosine Modulated Model)

2. Sinusoids - Sinusoidal waveforms occurring in the frequency region defined by the pitch-period. Experimental results have shown that the family of sinusoidal wave-functions only occur in that frequency region defined by the pitch-period. Higher frequency regions contain waveforms with Gaussian envelopes. The occurrence of a sinusoid is functionally dependent on the pitch-period of the voice. As the male voice pitch-period shortens, the wave-function structure changes from a Gaussian envelope character to a sinusoidal character. This is illustrated in Figure C-1a and b for a male speaker uttering the vowel /i/ as in "eve" at pitch-periods of 7.8 msec and 4.1 msec respectively. As the pitch-period shortens, the wave-function structure changes from a Gaussian to a sinusoidal characteristic. Two additional examples illustrate this effect. Since the female voice typically has a short pitch-period relative to the male voice, it would be expected from the above that the female voice would have a strong sinusoidal component for the voiced sounds. Figure C-2 shows a comparison between a normal male and female voice uttering the word "put". The male wave-function structure (Figure C-2a) has a consistent Gaussian characteristic for both the plosive and vowel sounds whereas the female voice exhibits a sinusoidal characteristic during the vowel segment. Figure 3 compares a normal male and female voice uttering the word "mat". The male voice (Figure 3a) exhibits a consistent Gaussian characteristic whereas the female voice (Figure 3b) shows a sinusoidal characteristic for the voiced nasal consonant and vowel, and then becomes Gaussian during the plosive.

The above examples show that a general wave-function analysis system must

1a



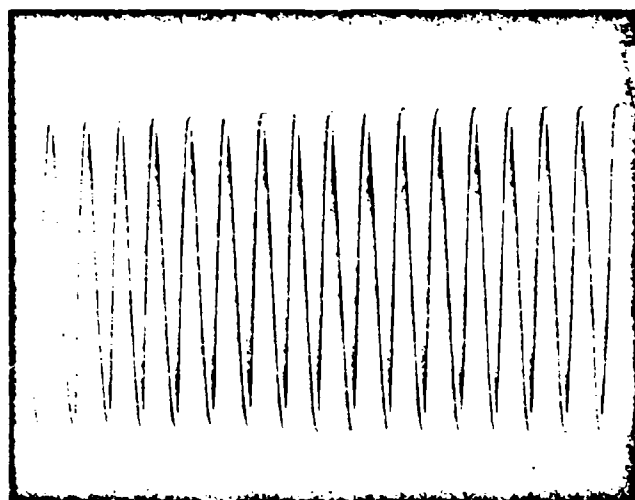
70 msec

/1/ eve

100 - 400 Hz

Pitch-period =
7.8 msec.

1b



70 msec

/1/ eve

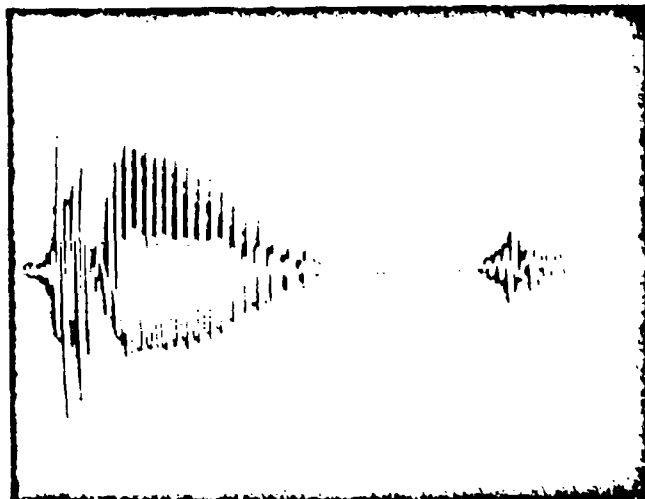
100 - 400 Hz

Pitch-period =
4.1 msec

Figure C-1 Male Speaker; Effect of Fitch on Wave-Function Structure

NOT REPRODUCIBLE

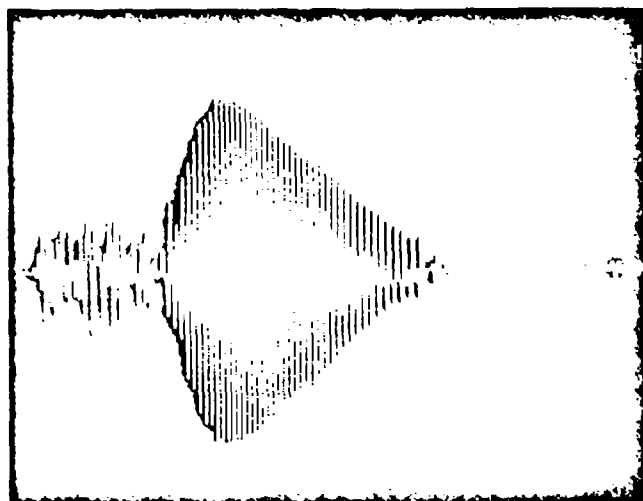
2a



MALE
100 - 400 Hz
"put"

420 msec.

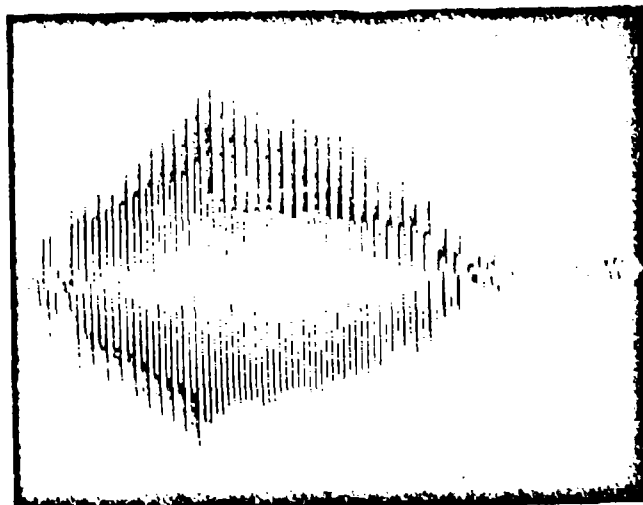
2b



FEMALE
100 - 400 Hz
"put"

Figure C-2 MALE vs. FEMALE Voice: Effect of Pitch on Wave-Function Structure

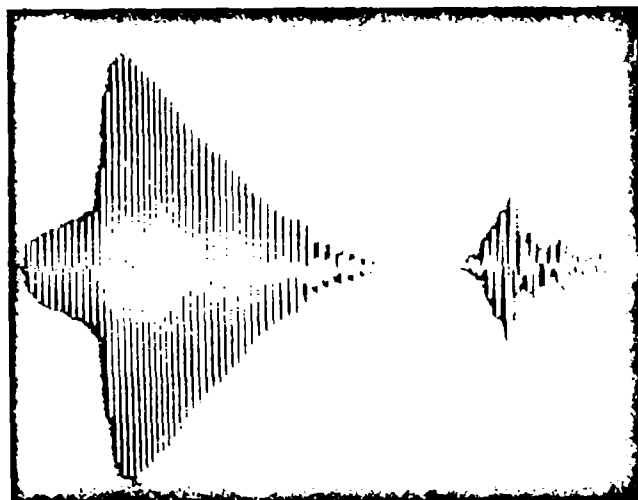
3a



MALE
100 - 400 Hz
"mat"

420 msec

3b



FEMALE
100 - 400 Hz
"mat"

420 msec

Figure C-3 Male vs. Female Voice; Effect of Pitch on Wave-Function Structure

be able to accurately represent wave-functions with Gaussian envelope characteristics and sinusoidal characteristics.

A method has been developed as part of the new analysis system that builds sinusoids out of GCM waveforms. This permits a consistent set of parameters to be generated by the new wave-function analyzer.

(b) Preprocessing of Acoustic Waveform

In the previous semi-annual report, a method was presented for filtering the input speech string into sub-strings that are amenable to wave-function analysis. This technique required the tracking of the major energy peaks in the short-term energy spectrum of the acoustic waveform (i.e. formant tracking for vowels) and adjusting the center frequency and bandwidths of these filters in covering the frequency range from 300 - 3200 Hz. It was demonstrated that this approach would correctly filter the original speech string.

This method of preprocessing the input speech data was employed successfully in both the analysis/synthesis and recognition studies. However, it became increasingly evident that the digital simulation of an automatic tracking filter was a complicated process and required the major portion of computer time in both the wave-function analysis/synthesis and recognition studies. In view of this further studies of the preprocessing problem were undertaken with the intention of defining a set of fixed frequency ranges which would correctly filter the acoustic waveform into acceptable sub-strings for both the male and female voice. This study has been successful in defining the four contiguous frequency bands

100 - 400 Hz	Band 1
400 - 900 Hz	Band 2

900 - 1800 Hz Band 3

1800 - 3600 Hz Band 4

which correctly partition the original speech string into four sub-strings whose waveform structure is of a form suitable for Gaussian wave-function analysis.

An experimental approach was taken in the determination of the fixed filter bandwidths and center frequencies. Initially the following criteria were established as the basis for the selection of the filter specifications:

1. Filter parameters must be selected so that the resulting sub-strings have an appropriate wave-function structure to fit the wave-function analysis model.
2. The number of fixed filter bands must be kept to a minimum to avoid additional complexity in the analysis scheme.
3. The fixed filter parameters must be chosen such that they contain relevant information for recognition.

The experimental endeavor involved generation of the short term energy spectrum of each of the 34 English phonemes for a male and female speaker. The first three major energy peaks for each phoneme were then plotted as a function of frequency. Examination of the resultant plots indicated that the energy peaks roughly occurred in three distinct frequency regions; below 1000Hz., 1000 - 2000 Hz., above 2000 Hz. These three fixed regions gave adequate filtering results even though two major energy peaks would be grouped together as for example for the vowel /a/. Further studies showed that breaking the region below 1000 Hz into two bands at 400 Hz and setting the upper frequency limit above 2000 Hz to 3600 Hz gave consistently good results. A frequency of 100 Hz was established as the lower frequency limit

to minimize random low amplitude noise occurring in the region below 100 Hz. The upper limit on Band 1 of 400 Hz was selected to correspond to a minimum pitch-period of 2.5 milliseconds. For the spectrum of pitch-periods encountered in the male and female voice this represents a reasonable choice. Additional experimental studies demonstrated that setting the upper limit on Band 2 at 900 Hz improved the results even more since this separated the second major energy peak from the first for several of the phonemes.

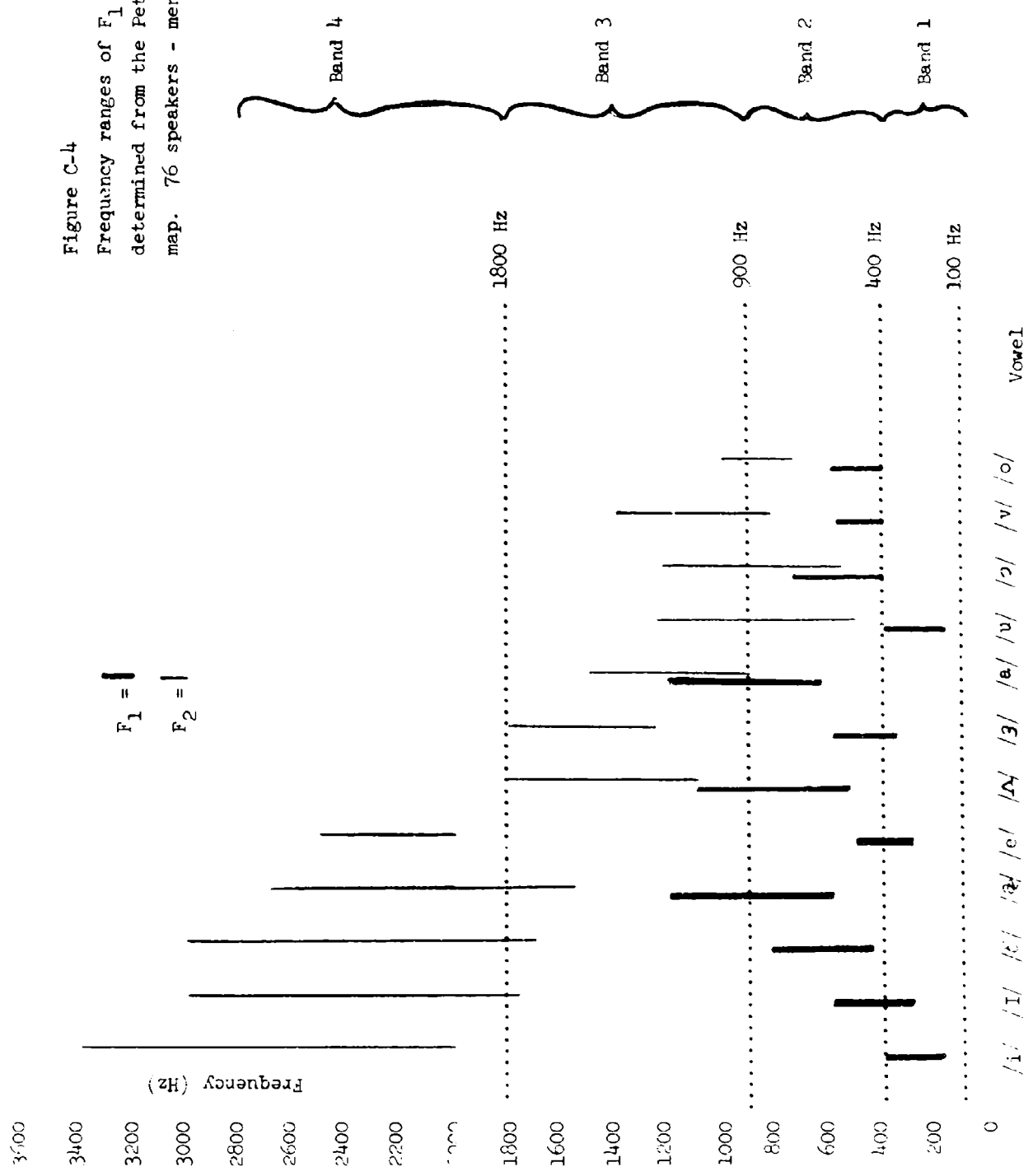
In the determination of the appropriate fixed filter bands it was useful to repeat the Peterson-Barney vowel map⁽¹⁾ into the line form depicted in Figure C-4. This vertical line plot represents the possible frequency ranges of the first two formants F_1 and F_2 (first two energy peaks) of the twelve vowels for a mix of 76 male, female, and child speakers.

The horizontal lines on Figure C-4 define the bandwidths of the four fixed filters. The figure illustrates that the frequency ranges defined by the four fixed filter bands do contain useful frequency information. From a consideration of the figure it can be seen that five vowels, (/i/, /I/, /e/, /ɛ/, /u/) have the possibility of the first formant occurring within Band 2, and in some cases the second formant may also occur within Band 2. Eight of the vowels can have a first or second formant occurring within Band 3, and seven vowels may have the second formant falling within Band 4. The point to note is that due to the positions of the formants within these fixed bands, some recognition information is available by a simple examination of the sub-string. For example, due to the lack of a formant in Band 3, the vowels /i/, /I/, /ɛ/, and /e/ will have very low

⁽¹⁾Peterson, Gordon E. and Barney, Harold L., "Control Methods Used in a Study of the Vowels", J. Acoustical Society Am., Vol. 24, pp. 175 - 184, March 1952.

Figure C-4

Frequency ranges of F_1 and F_2 as determined from the Peterson-Barney vowel map. 76 speakers - men, women, and children



amplitude in this band as compared to the other three bands. This is clearly illustrated in Figure C-5 which shows the output of the fixed filter pre-processor for the word "steek". Figure C-5a shows the word prior to filtering and Figure C-5b illustrates the four filtered sub-strings. In the 900 - 1800 Hz sub-string it can be seen that during the vowel portion (/i/) the amplitude is indeed almost insignificant as compared to the other sub-strings. Also note that the largest amplitudes occur in Bands 1 and 4, because these are the bands in which the formants occur.

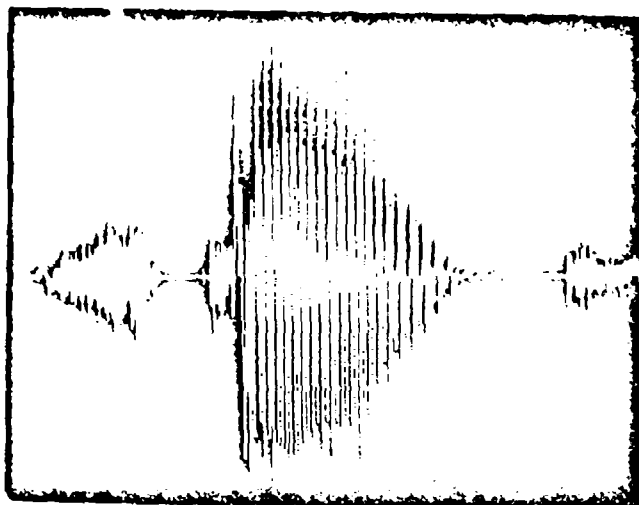
Consider the usefulness of these four frequency bands with respect to phonemes other than vowels. Figure C-6 shows the word "shop" and its four sub-strings. Note that the fricative phoneme /sh/ stands out clearly in the 1800 - 3600 Hz sub-string, the vowel portion stands out as a repetitive structure in all four sub-strings, and the /p/ phoneme is indicated primarily as a burst of low frequency wave-functions in the 100 - 400 Hz. sub-string. The combination of bands 1 and 4 can serve as strong indicators of voiced vs. unvoiced phonemes.

Another example is the word "men" as shown in Figure C-7. This example demonstrates how a nasal phoneme, with its voiced repetitive-like structure, can be distinguished from a vowel. Most of the power of a nasal resides in the 100 - 400 Hz. band while the vowel has significant energy in at least three bands.

The fixed filter ranges, although experimentally determined, do exhibit the common property that there is approximately an octave change across the filter bandwidth. For example the lower frequency of Band 3 is 900 Hz while the upper frequency is 1800 Hz., an octave difference. Note that Band 1 is

"STEAK"

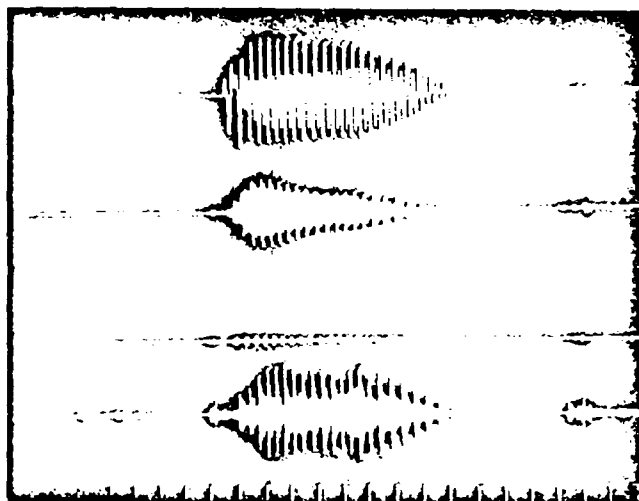
C-5a



Unfiltered

420 ms.

C-5b



100 - 400 Hz

400 - 900 Hz

900 - 1800 Hz

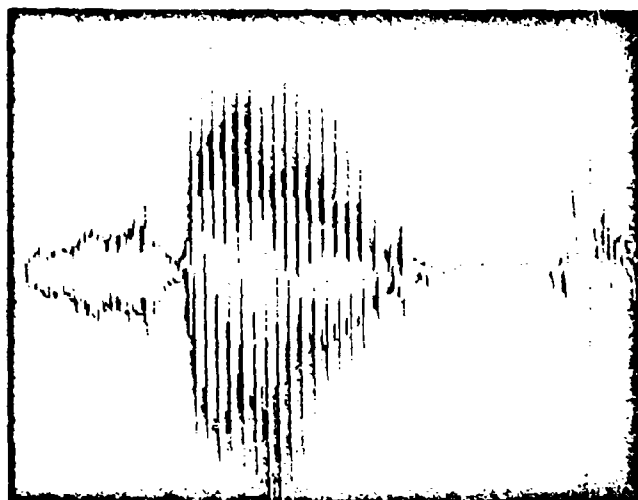
1800 - 3600 Hz

All $\frac{\sin x}{x}$ filter
kernels

Figure C-5 (a) The word "steak", unfiltered
(b) The 4 sub-strings of the word "steak"

"SHOP"

C-6a



Unfiltered

420 ms.

C-6b



100 - 400 Hz.

400 - 900 Hz.

900 - 1800 Hz.

1800 - 3600 Hz.

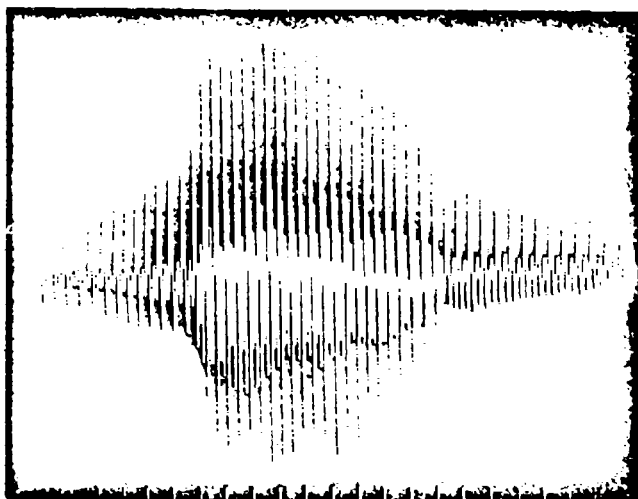
 $\frac{\sin x}{x}$ kernels

Figure C-6 The word "shop" and its four sub-strings.

NOT REPRODUCIBLE

"MEN"

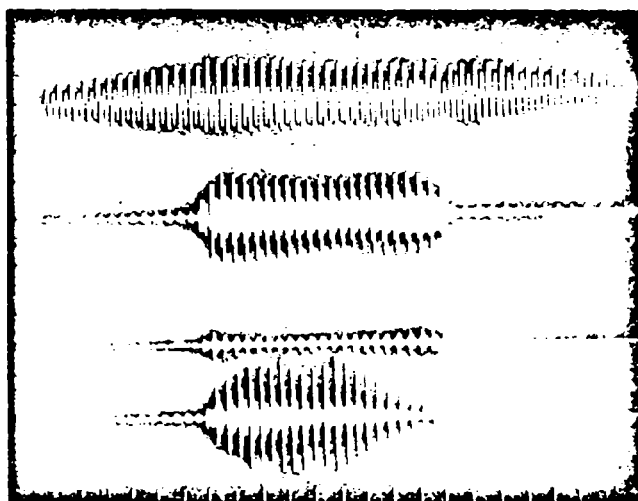
C-7a



Unfiltered

420 ms.

C-7b



100 - 400 Hz

400 - 900 Hz.

900 - 1800 Hz.

1800 - 3600 Hz.

All $\frac{\sin x}{x}$ filter

kernels

Figure C-7 The word "men" and its four sub-strings.

the one exception to this observation.

Figure C-8 compares the original and synthetic versions of the word "max" as recorded for a male speaker for the adaptive and fixed filtering approaches. As indicated the synthesized version of "max" using adaptive filtering (Figure C-8b) and that using fixed filtering (Figure C-8c) compare favorably with the original word.

The purpose of the preprocessor is to transform the input acoustic waveform denoted as a "string" into "sub-strings" that are amenable to wave-function analysis. It has been demonstrated above that a high-quality wave-function representation can be obtained by filtering the input string into four sub-strings covering the frequency range from (100, 3600) Hz. Let $s(t)$ be the original string with frequency components in the range (100,3600)Hz. Then

$$s(t) = \sum_{n=1}^4 s_n(t) \quad (C-1)$$

where $s_n(t)$ is the n^{th} sub-string. In the frequency domain

$$S(j\omega) = \sum_{n=1}^4 S_n(j\omega)$$

The frequency regions R_n corresponding to $S_n(j\omega)$, $n = 1, 4$ are divided in the following manner:

$$100 < R_1 < 400 \text{ Hz}$$

$$400 < R_2 < 900 \text{ Hz}$$

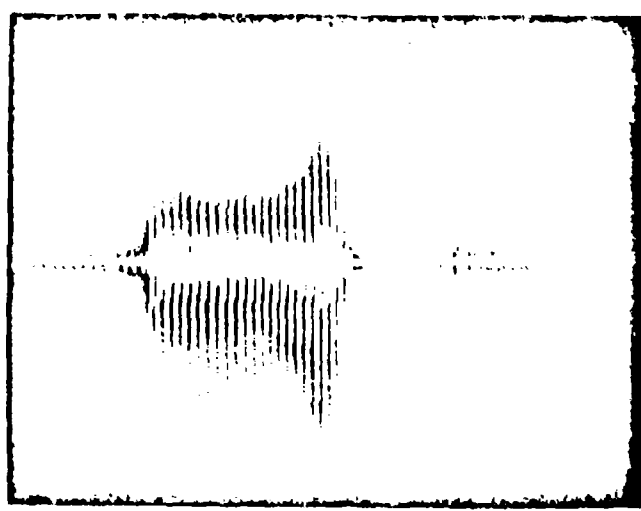
$$900 < R_3 < 1700 \text{ Hz}$$

$$1700 < R_4 < 3600 \text{ Hz}$$

This separation into four contiguous frequency regions corresponds to convolution of $\sin x/x$ type bandpass filters with $s(t)$ to obtain the sub-strings.

NOT REPRODUCIBLE

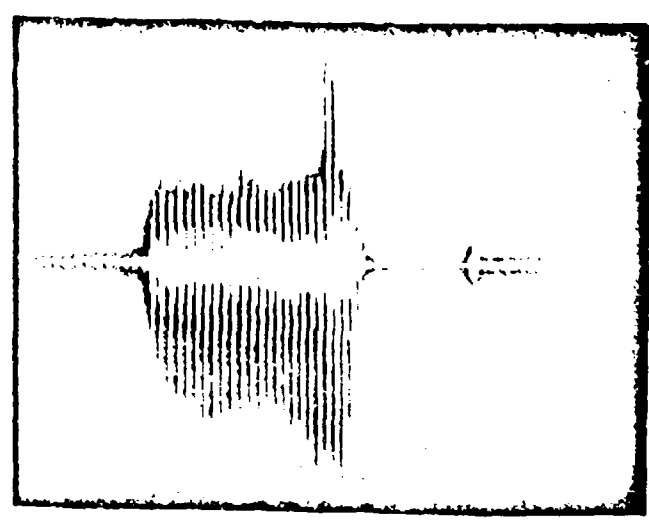
C-8a



Original

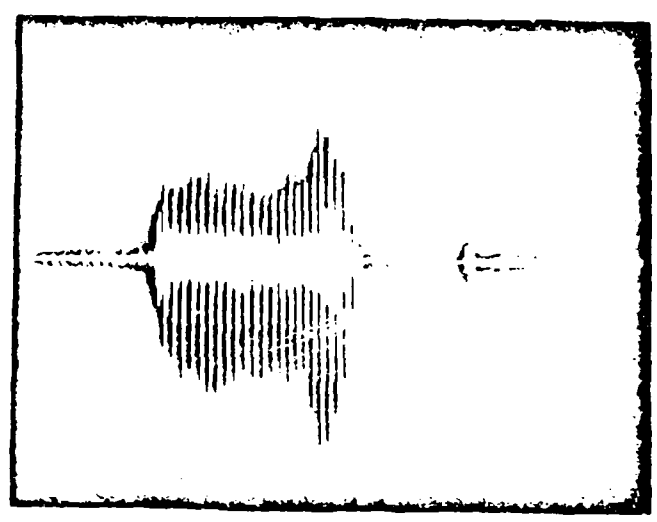
420 ms

C-8b



Synthetic
Adaptive Filter

C-8c



Synthetic
Fixed Filter

Figure C-8 Comparison of original and synthetic versions of word "max" from male speaker using adaptive and fixed filtering.

By applying the discrete convolution equation each of the four sub-strings for the system simulation is obtained as

$$s_n(k) = \sum_{j=-62}^{62} s(k) h_n(j-k) \quad (C-2)$$

where $s_n(k) = S_n(t)|_{t=kT}$ $k = 0, 1, 2, \dots$

$$h_n(j-k) = 2B_n \frac{\sin(2\pi B_n(j-k))}{2\pi B_n(j-k)} \cos[2\pi F_j(j-k)]$$

and T is the discrete sampling period. The parameters of the n^{th} convolution kernel $h_n(t)$ are defined from the n^{th} region. For example $B_2 = 900 - 400 = 500$ Hz and $F_2 = (400 + 900)/2 = 650$ Hz. Equation (C-2) is utilized for simulating the fixed-filter preprocessor on the UCSB 1800 speech system.

(c) Improved Wave-Function Analysis/Synthesis System

The process for wave-function analysis of human speech that has evolved at UCSB is a three-step operation.

1. Record a sample of speech of time length T
2. Preprocess (filter) the speech sample into four sub-strings each of duration T
3. Analyze each sub-string into its set of ASC ϕ N parameters.

Previous work at UCSB has used the Gaussian Wave-Function family as a model for the individual wave-functions in the filtered sub-strings of the raw speech string. The family of Gaussian wave-functions is the set of derivatives of the Gaussian function $e^{-t^2/2}$. The n^{th} function is explicitly described by the Gaussian function multiplied by a Hermite polynomial of degree n . The family of functions satisfy the differential equation

$$0 = \ddot{U}(t) + \left(\frac{2\pi}{s}\right)^2 (t-c)\dot{U}(t) + \left(\frac{2\pi}{s}\right)^2 (N^2 - \frac{1}{2})U(t) \quad (C-3)$$

with initial conditions

$$\begin{aligned} U(C) &= A \cos \phi \\ \dot{U}(C) &= A \frac{2\pi N}{S} \sin \phi \end{aligned}$$

N is a physically descriptive parameter defining the number of half cycles of the Hermite Polynomial which occur under the envelope of a particular wave-function. The relationship defining N is

$$N = \sqrt{n + 3/2}$$

where n = order of the Hermite Polynomial. There is no closed form solution which describes this family of wave-functions. It has been necessary to implement a recursive solution to Equation (C-3) in order to generate any arbitrary wave-function of this family. This complicates the problem of analysis and synthesis.

As reported previously, an asymptotic solution to Equation (C-3) has been obtained which is of a closed form and is also valid for the range of wave-functions encountered in human speech. This solution defines the Gaussian Cosine Modulated (GCM) family of wave-functions. Any arbitrary member of this family of wave-functions can be described by

$$U(t) = Ae^{-\left(\frac{\pi}{S} (t-C)\right)^2} \cos (\omega_0 (t-C) - \phi) \quad (C-4)$$

$$\omega_0 = 2\pi F_0$$

The solution is thus a Gaussian envelope of amplitude A , center in time of C , and spread in time of S , multiplied by a cosine wave of frequency F_0 and phase ϕ with respect to C .

This model still has only five parameters describing the entire wave-function but now the set of ASC ϕ N parameters is replaced by the set of AFC ϕ F

parameters. A representative wave-function is shown in Figure C-9.

The parameters of the GCM model as in the Gaussian Wave-Function Model are chosen to be physically meaningful in describing the given wave-function.

A = Amplitude of envelope of wave-function

S = Spread of wave-function envelope or that time interval during which 99 % of the energy of the wave-function occurs.

C = Center of the envelope in time

ϕ = Phase of the cosine wave with respect to C

F = Frequency of the cosine wave

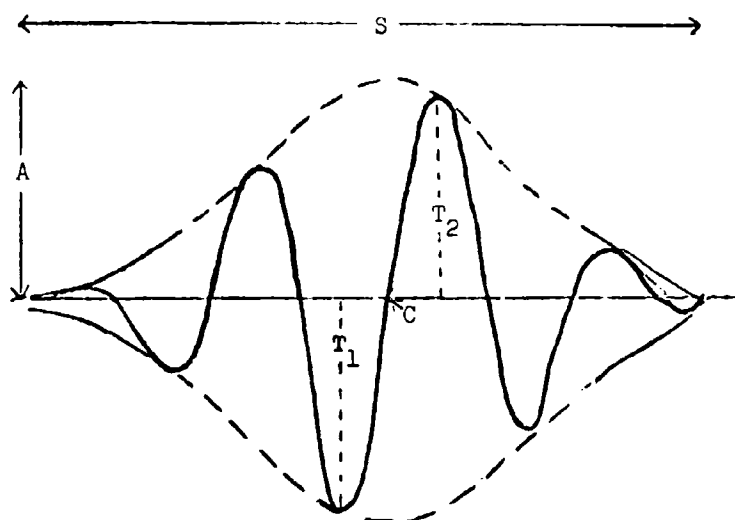
From a computational and conceptual viewpoint the GCM model is much simpler to operate with. Therefore a new analysis and synthesis has been developed based upon this model.

GCM Wave-Function Analysis

Previous work in developing the wave-function analysis process required a multi-pass analysis on a given sub-string of data. This process is undesirable because it generates multiple sets of wave-function parameters for each sub-string and multiplies the amount of time required to analyze a given sub-string by the number of passes on the sub-string. Investigation showed that multi-pass analysis was required due to three problem areas.

- 1) Improper preprocessing (filtering) of the raw speech string
- 2) Inaccuracies in the analysis process
 - a) Sampled data inaccuracies
 - b) Failure to set practical limits on calculated parameter values
- 3) Inability to handle sinusoids, particularly on the female voice.

From an ideal standpoint, a one pass analysis system is desirable in order



$$C = \frac{T_2 + T_1}{2}$$

$$\phi = 90^\circ$$

$$F_0 = \frac{1}{2(T_2 - T_1)}$$

Figure C-9

Representative GCM Wave-Function

to achieve maximum speed in the analysis process and a minimum number of parameters to describe the speech sub-string.

In order to achieve this goal, a new analysis system, based upon the GCM model, has been developed which accurately performs a one pass analysis on any arbitrary speech input from a male or female voice. In conjunction with this, a GCM based wave-function speech synthesis system has also been completed.

Each of the problem areas that necessitated multi-pass analysis has been investigated and the appropriate solution has been implemented in the new system. Problem area one, improper preprocessing, has been solved by the definition of the four correct fixed filter bands previously discussed and then constructing the appropriate $\frac{\sin x}{x}$ band pass kernels to use in the existing digital filter convolution programs. Problem area two, inaccurate analysis, has been solved by performing an error analysis on the wave-function process to define the significant analysis errors that needed correction.

These were:

- 1) Improper estimation of extrema and times of occurrence due to inaccuracies in the sampled data. This factor introduces a significant error, that is a function of frequency, into all five parameters. The error is minimized by implementing a parabolic curve fitting operation to the sampled data during the extrema detection operation.

- 2) Failure to set bounds on calculated parameters. The sampled speech data varies at times significantly enough from the wave-function model, so that, unless the S parameter is bounded, irrecoverable errors are introduced in the Residue calculation process. To avoid this, when analyzing a given wave-function, the speech string is extrapolated into the future to locate the

next wave-function. S is then bounded so the present wave-function does not couple beyond the center C of the future wave-function. The character of the future wave-function is therefore not destroyed by an error in the calculation of the parameters for the present wave-function. Errors in the other four parameters were found to be generally small enough to avoid the necessity of bounding them.

The third problem area, inability to handle sinusoidal wave-functions, was solved by

- 1) Setting filter Band 1 to an upper limit of 400 Hz. This limits the sinusoidal component to Band 1.

- 2) Construction an algorithm in the analysis process that detects the presence of a sinusoid and then generates a set of ASCOF parameters from which the sinusoid can be built. The algorithm is specifically tailored to generate one wave-function per pitch period, even during a sinusoid, so that the pitch information contained in the C parameter is retained for the recognition system.

As in previous work, the new wave-function analysis system is based upon a four point analysis of a given wave-function that uses the four extrema grouped around C to calculate the five parameters to define the wave-function. This requires an extrema detection process which maps the sampled data format into an extrema (peak vs. time) format. Therefore the first step prior to the analysis process on the filtered sub-string is to convert the sampled data sub-string into a list of extrema. Define

$S(t)$ = Acoustic waveform speech string

$S_d(t)$ = Sampled acoustic waveform speech string

$S_{dn}(t)$ = Sampled filtered sub-string $n = 1, 2, 3, 4$

ω_{Kn} = Extrema listing of speech string $n = 1, 2, 3, 4$

A block diagram of the process prior to analysis would then be as shown in Figure C-10.

The analysis process then consists of scanning the extrema list ω_{Kn} to isolate the wave-functions in the sub-string and then to calculate the five ASCOF parameters which define each isolated wave-function. Since four points are used to characterize a wave-function, the extrema list is scanned using four points at a time until a stopping condition for a wave-function occurs. Determining the ASCOF parameters is therefore a two step process: 1) Satisfy stopping criteria to isolate a wave-function 2) Calculate ASCOF parameters. Once the parameters of a given wave-function have been determined, the effect of the wave-function coupling into the future must be removed to be able to correctly determine the parameters of the next wave-function. This is accomplished by building the calculated wave-function and then subtracting out its effect from the extrema listing. This process is the Residue calculation.

A block diagram of the entire analysis process is shown in Figure C-11.

Extrema Detection and Correction

The sub-string to be analyzed is in a sampled data form. This is converted to a sign magnitude (peak) vs. time format by the extremum detection and correction operation.

Let

$x_j = j^{\text{th}}$ sample data point

$j = \text{index of sample data list } j = 1, \dots, 7440$

The operation is defined in the flow chart of Figure C-12. The analysis system used depends upon accurate extrema parameter values. Since sampled data is

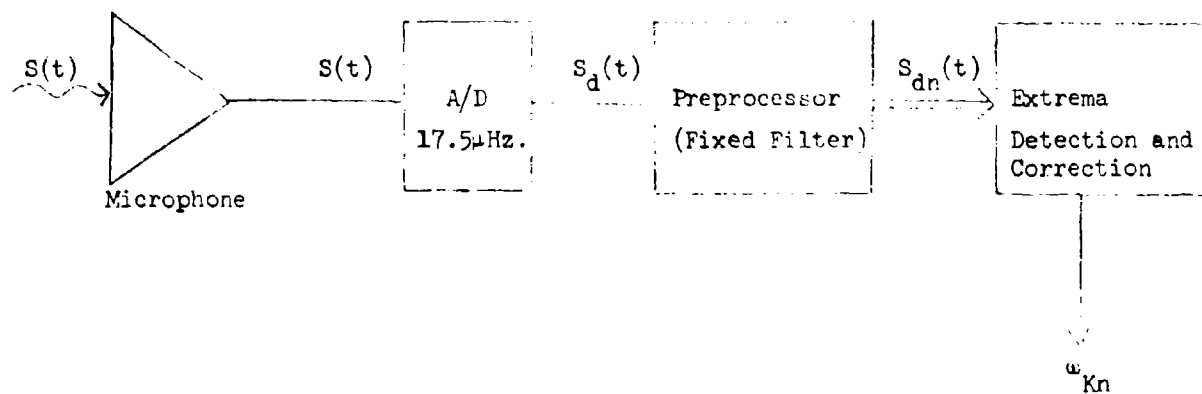
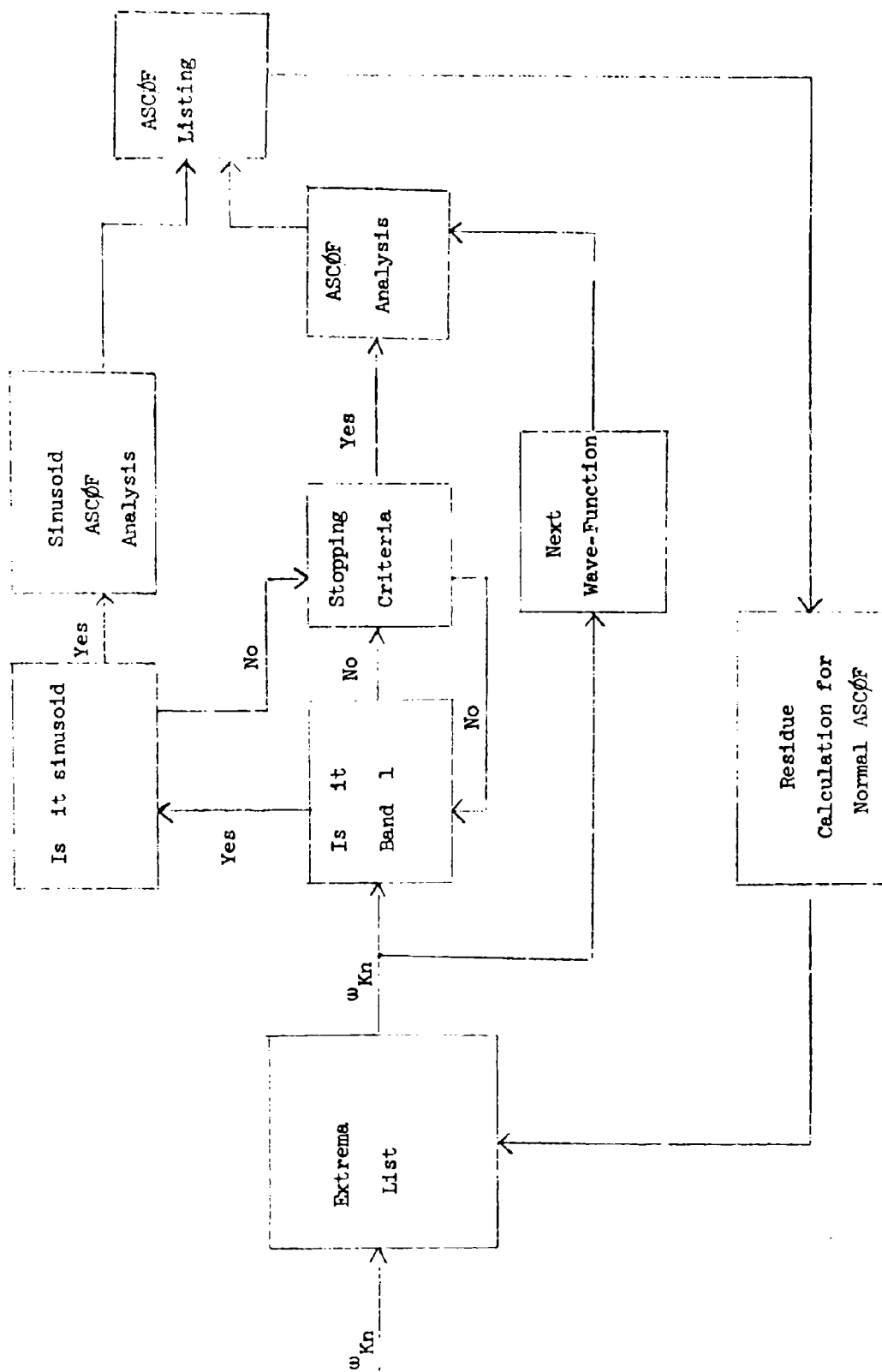


Figure C-10 Block diagram of preprocessing and extrema conversion processes.

Figure C-11 Block diagram of wave-function analysis process



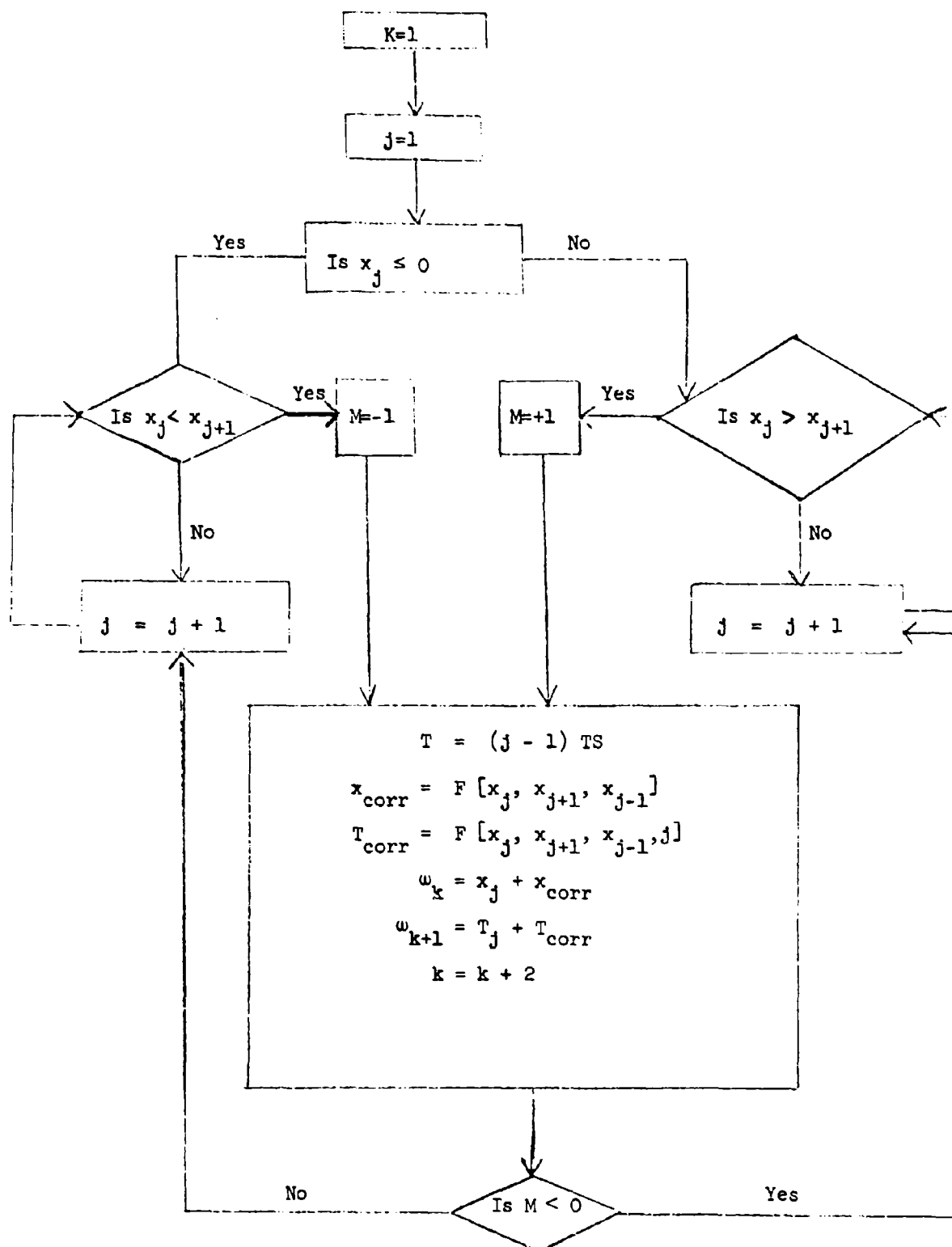


Figure C-12 Extrema detection and correction flow-chart

only an approximation to the true extrema, a parabolic curve fitting is done to the sampled data to precisely define the extrema. This is accomplished by the following relations.

Sampling Frequency = 17.5 kHz

$TS = 1/17.5$

$T = (j - 1) \times TS = \text{Time in msec. of } j^{\text{th}} \text{ sample}$

$A = (x_{j+1} - 2x_j + x_{j-1}) / (2 TS^2)$

$B = (x_{j+1} - x_{j-1}) / (2 TS)$

$X_{\text{CORR}} = -B^2 / (4A) = \text{Correction to peak value}$

$T_{\text{CORR}} = -B / (2A) = \text{Correction to time value}$

(C-5)

Stopping Criterion (Normal)

When scanning the extrema data list ω_n , a criterion is established to isolate a wave-function behavior

$K = \text{Extrema data list index}$

$\omega = \text{Extrema data list}$

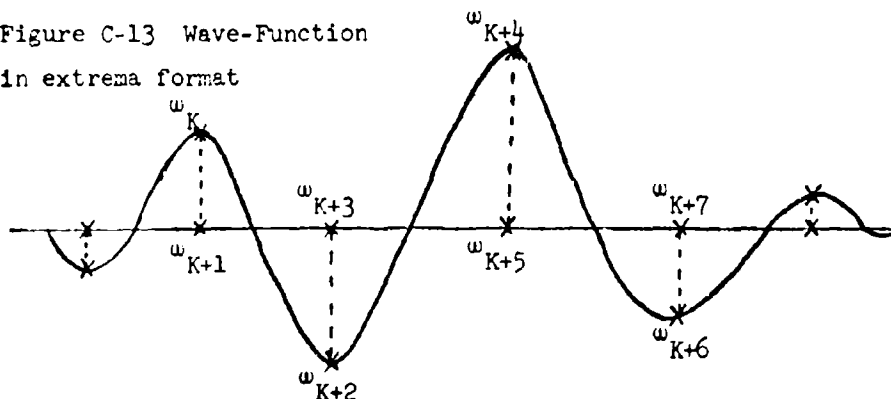
$K = 1 = \text{Initial condition}$

$\omega_K = \text{Sign magnitude of extrema}$

$\omega_{K+1} = \text{Time of occurrence of extrema}$

Figure C-13 illustrates the character of a wave-function in extrema format.

Figure C-13 Wave-Function
in extrema format



The following quantities can be defined from Figure C-13 for a wave-function in extrema format.

K = present index of extrema data list

ω_K = Peak 1

ω_{K+1} = Time of peak 1

ω_{K+2} = Peak 2

ω_{K+3} = Time of peak 2

ω_{K+4} = Peak 3

ω_{K+5} = Time of peak 3

ω_{K+6} = Peak 4

ω_{K+7} = Time of peak 4

The stopping criterion is

STOP if

$$|\omega_{K+2}| \geq |\omega_{K+6}|$$

and

$$|\omega_{K+4}| \geq |\omega_K|$$

otherwise increment the index by 2 to $K + 2$ and check for the stopping criterion again.

If the stopping criterion is satisfied, this means that A and C will occur during the time interval defined by $(\omega_{K+3}, \omega_{K+5})$. The stopping criterion is equivalent to the presence of a local maximum or minimum in the extrema data list.

Once the normal stopping criterion has been satisfied, the ASCOF parameters of the isolated wave-function are then calculated. This process is accomplished by making calculations based on the known geometry of the wave-function family.

The four extrema define four points in the envelope of the wave-function. This is sufficient information to calculate the S and C parameters. Once S and C are known, A and ϕ can be calculated from either of the extrema adjacent to C. The time values of the two extrema around C can be used to determine the frequency parameter. The ASCOF parameters are thus found from the following expressions.

Determination of F

ω_{K+5} = Time value of extrema to right of C

ω_{K+3} = Time value of extrema to left of C

$$F = \frac{1}{2(\omega_{K+5} - \omega_{K+3})} \quad (C-6)$$

Choosing the two time values at the center of the wave-function minimizes errors in the frequency calculation due to coupling from adjacent wave-functions.

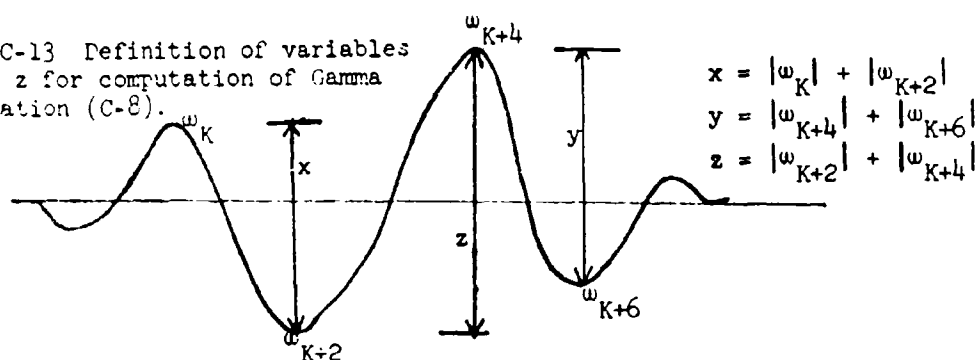
Calculation of S

$$S = \frac{\pi}{F} \left[\frac{-1}{2 \ln [2(\text{Gamma})]} \right]^{\frac{1}{2}} \quad (C-7)$$

$$\left. \begin{aligned} \text{Gamma} &= \frac{[|\omega_K| + |\omega_{K+2}|] [|\omega_{K+4}| + |\omega_{K+6}|]}{[|\omega_{K+2}| + |\omega_{K+4}|]^2} \\ \text{Gamma} &= \frac{(x)(y)}{(z)^2} \end{aligned} \right\} \quad (C-8)$$

where these variables are defined in Figure C-13.

Figure C-13 Definition of variables x, y, and z for computation of Gamma for Equation (C-8).



Computation of C

$$C = U \omega_{K+3} + (1 - U) \omega_{K+5}$$

where

$$\left. \begin{aligned} \text{RHO} &= \frac{|\omega_{K+y}| - |\omega_K|}{|\omega_{K+2}| - |\omega_{K+6}|} \\ U &= [\pi/2 - \text{atan}(\text{RHO})] 2/\pi \end{aligned} \right\} \begin{array}{l} x > y \\ \\ \end{array} \quad (C-9)$$

$$\left. \begin{aligned} \text{RHO} &= \frac{|\omega_{K+2}| - |\omega_{K+6}|}{|\omega_{K+y}| - |\omega_K|} \\ U &= \text{atan}(\text{RHO}) 2/\pi \end{aligned} \right\} \begin{array}{l} \\ \\ x < y \end{array}$$

Calculation of A

$$A = |\omega_{K+2}| e^{-[\pi/3(\omega_{K+3} - C)]^2} \quad (C-10)$$

Calculation of ϕ

$$\left. \begin{aligned} \phi &= 2\pi F (\omega_{K+3} - C) & \omega_{K+2} < 0 \\ \phi &= 2\pi F (\omega_{K+3} - C) + \pi & \omega_{K+2} > 0 \end{aligned} \right\} \quad (C-11)$$

Equations (C-5) through (C-11) represent the basic wave-function analysis method. To compensate for deviations in the speech data from the basic GCM model, two additions are required

- 1) Band 1 sinusoidal analysis
- 2) Upper limit on S

Sinusoid Analysis

The sinusoidal component takes the form of a sinusoid multiplied by the volume emphasis function of the human voice. As such the normal stopping criterion for a GCM wave-function will not detect the existence of a sinusoid. The transition into steady state portion, and transition out of the voiced sinusoid must be

detected. This is accomplished by again forming ratios based upon the four extrema being tested in the extrema list. The sinusoid is characterized by all four of the adjacent extrema being of approximately the same magnitude. Therefore the sinusoid stopping criterion is as follows using Band 1 information.

Is $x \leq z$

Yes $ETA = \frac{x}{z}$

No $ETA = z/x$

Is $ETA < .9$

Yes Go to GCM Stopping Criterion

No Is $y \leq z$

Yes $ETA = y/z$

No $ETA = z/y$

Is $ETA < .9$

Yes Go to GCM Stopping Criterion

No A sinusoid exists

Then set

$\Gamma = .3$

$U = .5$

and set these values into the normal set of equations to calculate the ASCOF parameters.

This will create a GCM wave-function to fill a pitch interval (one cycle of the sinusoid) of the voiced sound. This process is graphically illustrated in Figure C-14.

Since there is no coupling from the derived GCM wave-function into the future, the Residue calculation process is bypassed.

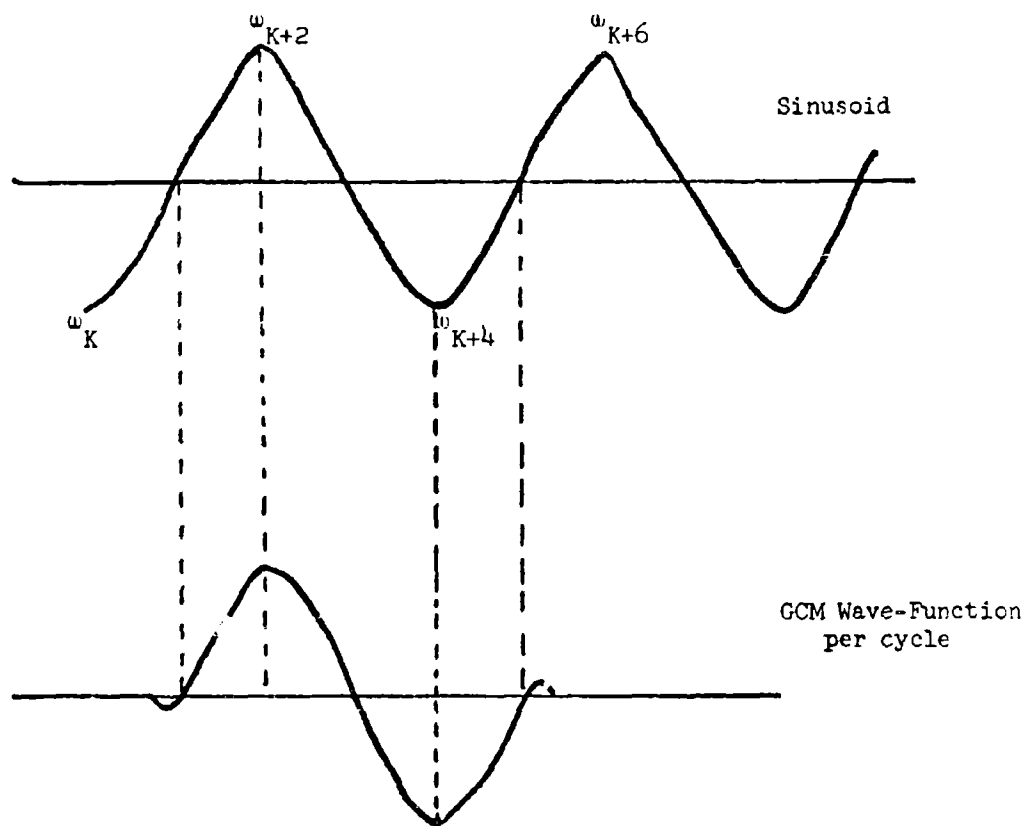
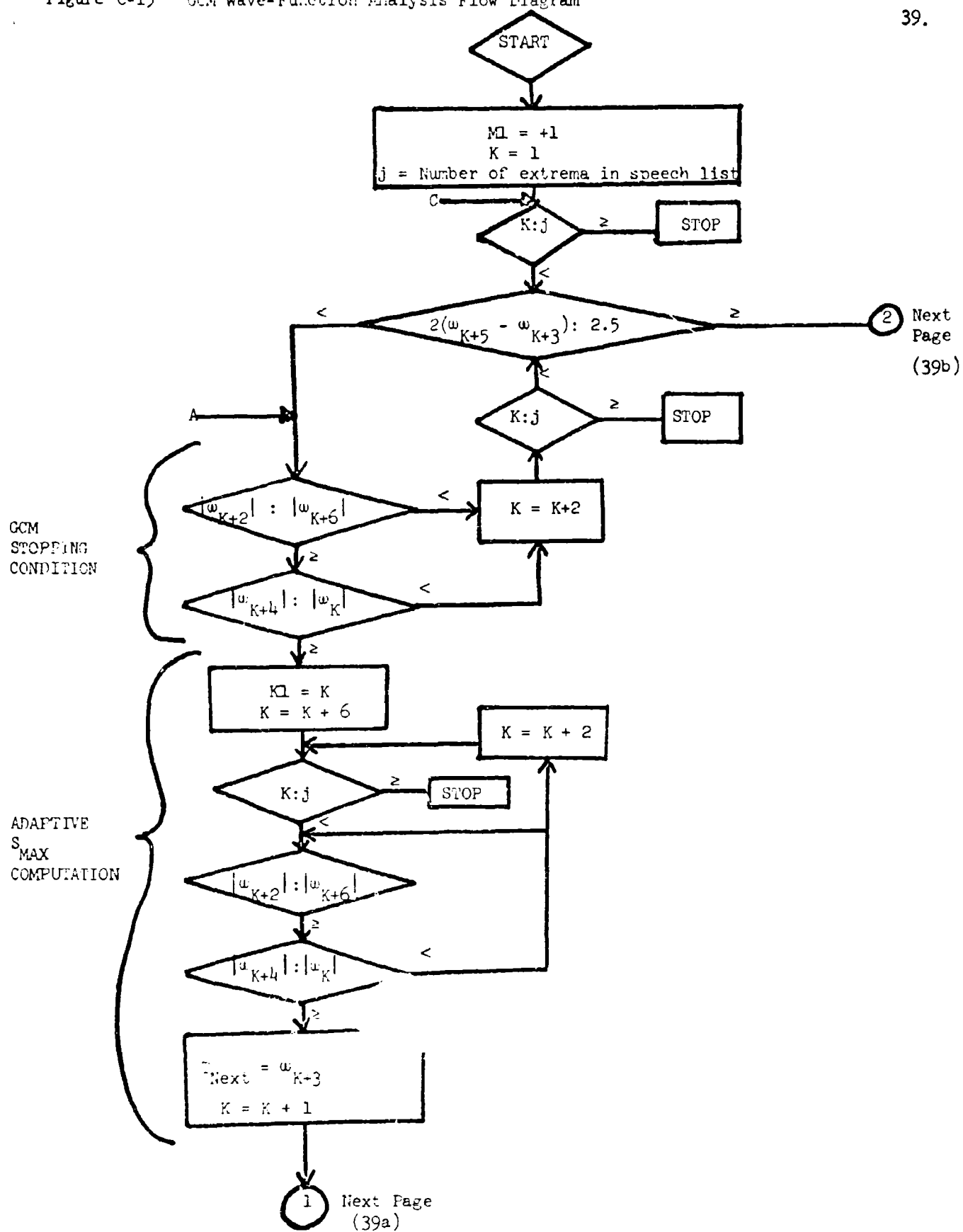
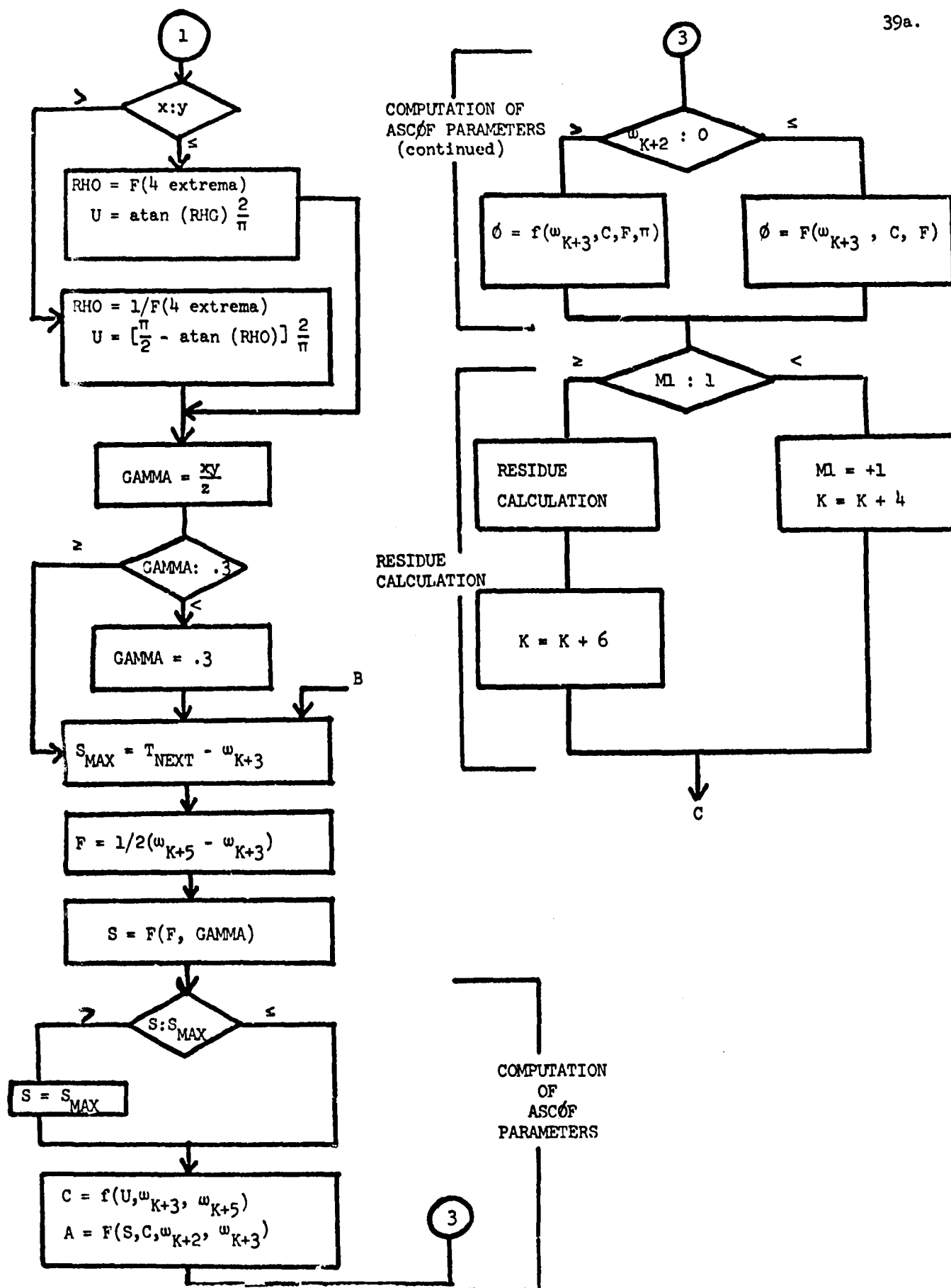
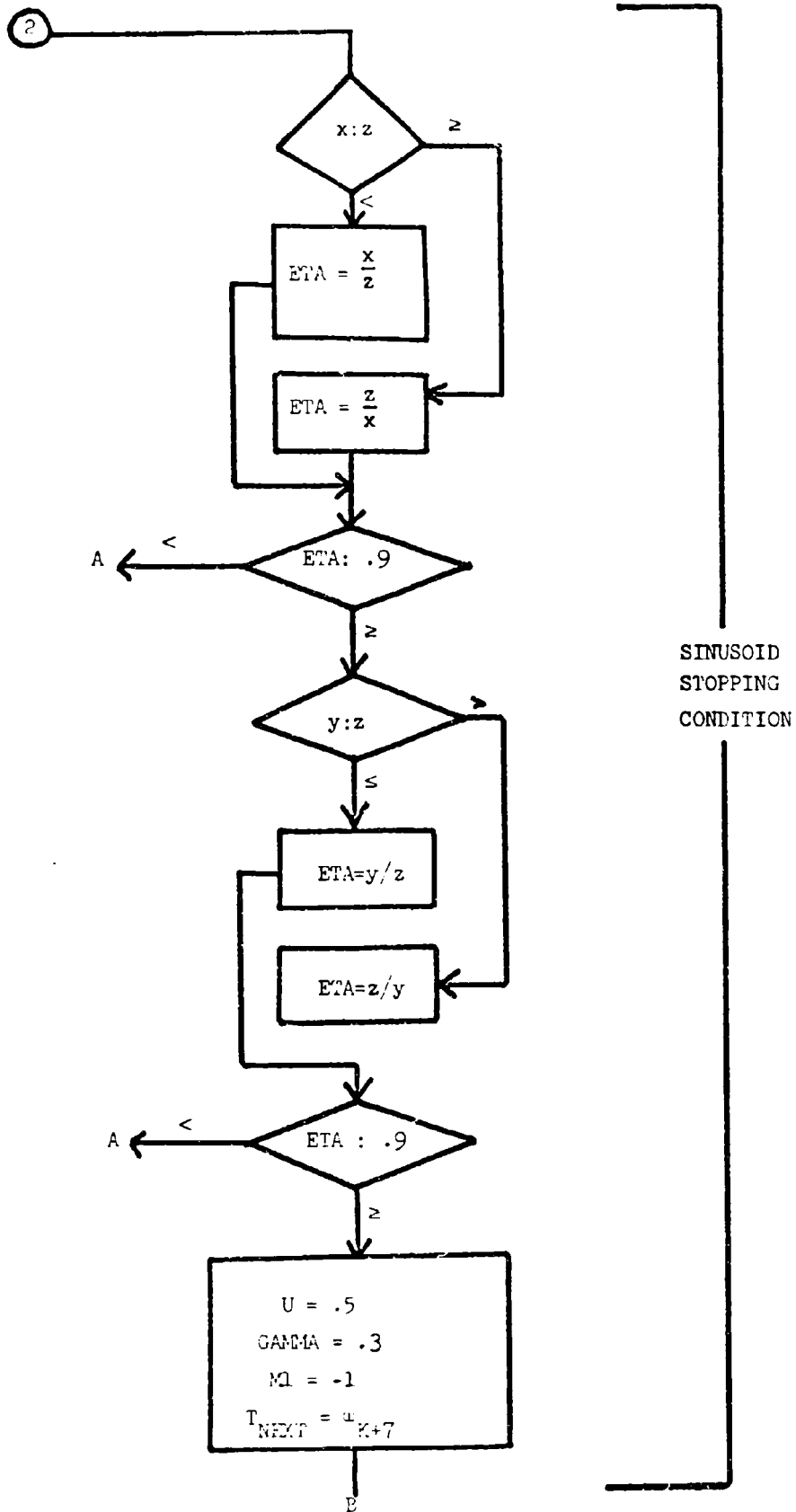


Figure C-14 Construction of one cycle of Sinusoid from a GCM Wave-Function

Figure C-15 GCM Wave-Function Analysis Flow Diagram







Upper Bound on S Parameter

Of the five parameters, errors in S have the most effect on introducing errors in the analysis process. This is because the S parameter defines the amount of coupling between adjacent wave-functions. There are significant enough deviations in the actual speech data from the GCM model to introduce unacceptable errors in the S calculation if S is left unbounded. To minimize error coupling between adjacent wave-functions S_{\max} is limited so that the present wave-function does not couple past the center of the next wave-function. The maximum value of S, S_{\max} , is given by the relation

$$S_{\max} = 2(C_{i+1} - C_i) \quad (C-12)$$

where i is the wave-function index.

S_{\max} can be approximated by

$$S_{\max} = 2[(\omega_{K+2})_{i+1} - (\omega_{K+2})_i] \quad (C-13)$$

The solution to Equation (C-13) is accomplished by searching the extrema list ahead in time until the next wave-function is isolated by the normal stopping criterion.

Lower Bound on Gamma

Gamma = .3 corresponds to four peaks existing under a Gaussian envelope in the GCM model. Since a four point analysis system is used $\text{Gamma}_{\min} = .3$. A flow-chart of the complete wave-function analysis process is shown in Figure C-15.

GCM Wave-Function Synthesis

Assuming that the ASCOF parameters have been correctly sorted into the four different sub-string sets, the synthesis is a straight forward process.

The synthesized sub-strings $\hat{s}_n(t)$, $n = 1, \dots, 4$ are obtained by application of Equation (C-4). Thus

$$\hat{s}_n(k) = \sum_i \varphi(\Omega(i,n),k) \quad (C-14)$$

where $\Omega(i,n)$ denotes the i^{th} set of wave-function parameters corresponding to the n^{th} sub-string, and k denotes the discrete time index; i.e.,

$$t = kT, \quad k = 0, 1, \dots \text{ where } T \text{ is the sampling period}$$

The symbol φ denotes the wave-function as a function of these variables.

In computational form,

$$\begin{aligned} \varphi(\Omega(i,n),k) = & A(i,n) \exp\left[[kT-C(i,n)]^2 \cdot \frac{\pi^2}{S(i,n)^2}\right] \cdot \\ & \cos\left[2\pi F(i,n) \cdot [kT-C(i,n)] - \phi(i,n)\right] \end{aligned} \quad (C-15)$$

where

$$\Omega(i,n) = [A(i,n), S(i,n), C(i,n), \phi(i,n), F(i,n)] \quad (C-16)$$

as the wave-function parameter set. Since each wave-function dies off as an exponential squared, only those located nearest to the corresponding present time t need to be evaluated at the index k .

The total synthesis of the estimated string $s(t)$ is denoted by $\hat{s}(t)$ and is calculated in discrete form by summing the sub-strings. Thus

$$\hat{s}(k) = \sum_{n=1}^4 \hat{s}_n(k) \quad (C-17)$$

Examples of the New GCM Wave-Function Analysis/Synthesis System

To illustrate the accuracy of the new, one pass, GCM wave-function analysis/synthesis system, representative phonemes and multiphoneme sounds from the male and female voice have been analyzed and synthesized. Figure C-17a shows a comparison between the original vs. synthetic waveforms of a 196 msec. segment

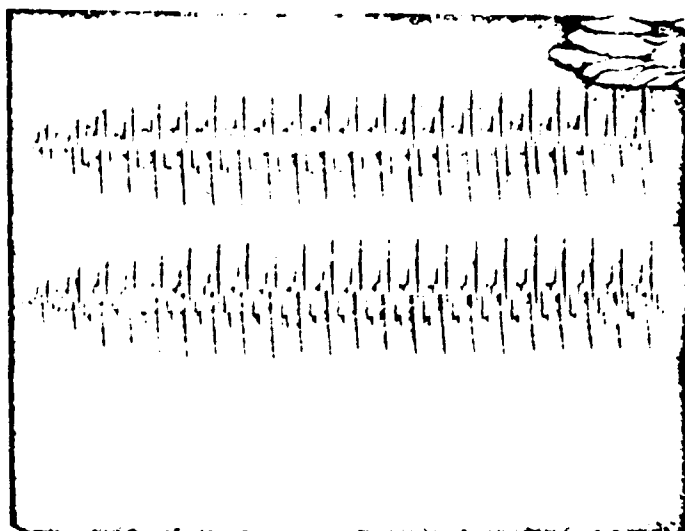
of Band 1 of the vowel /e/ as in "met". Figure C-17b is a more detailed view of a 35 msec. segment of the same vowel with the synthetic (dotted) waveform plotted over the original. A comparison of the original vs. synthetic waveforms of a 47 msec. segment of Band 2 of the same vowel is depicted in Figure 18a. Figure 18b shows a detailed 35 msec. comparison of the same vowel. In Figure 19a the synthetic and original waveforms for Band 2 of a 119 msec. segment of the vowel /a/ as in "all" are illustrated with Figure 19b showing a 35 msec. detailed comparison. Figure C-20 compares the synthetic and original waveforms for Band 3 of the same vowel. Note that here, the analyzer fitted in only one function per pitch period. Figure C-21 shows Band 1 of the synthetic and original waveforms of the fricative consonant /f/ as in "for". A comparison of Band 4 of the synthetic and original waveforms of the fricative consonant /sh/ as in "she" is illustrated in Figure C-22. Figure C-23 shows Band 4 of the synthetic and original waveforms of the stop consonant /t/ as in "to".

Looking at multi-phoneme sounds Figure C-24 shows Band 1 of the word "me" uttered by a female speaker. Both the nasal consonant /m/ and the vowel /i/ were sinusoidal and were both accurately represented. Figure C-25 illustrates the "ei" part of the word "pfeifer" uttered by a male speaker. This figure shows that even the coupling between vowels is accurately described by the analyzer. Figures C-26a through e show a detailed comparison of each of the four bands and the synthetic versus raw speech string for the word "pete" spoken by a female. Examining Figure C-26a which shows the four synthetic substrings summed together to form the synthetic speech string, demonstrates how accurately the synthetic speech duplicates the original.

The GCM analysis/synthesis system described in this section will be utilized

NOT REPRODUCIBLE

a)

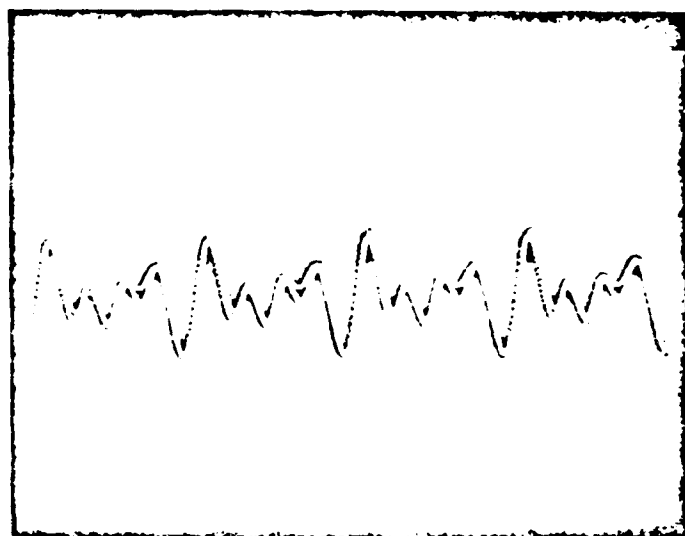


Synthetic

Original

196 msec. segment

b)



Solid = Original

Dotted = Synthetic

35 msec. segment

Figure C-17 Original versus synthetic waveforms of vowel /e/ as in "met"; Band 1.

a) 147 msec. segment

Synthetic

Original

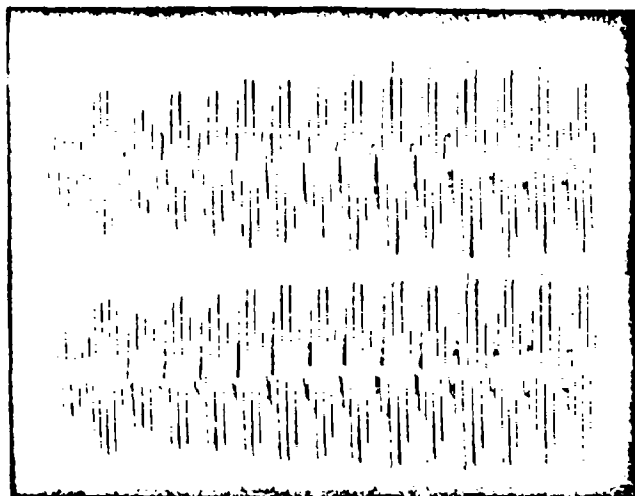
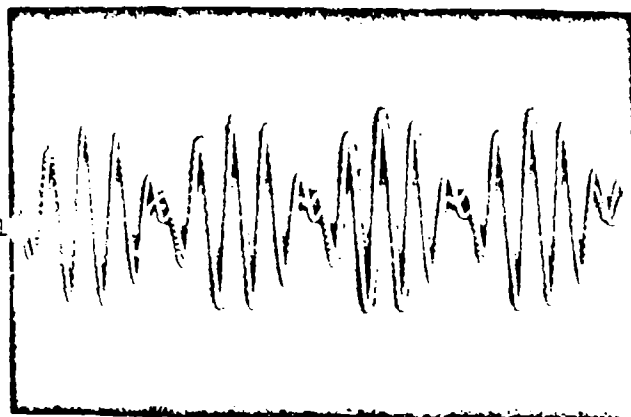


Figure C-18 Band 2 synthetic and original waveforms for vowel /ε/ as in "met". b) 35 msec. segment

Solid = Original

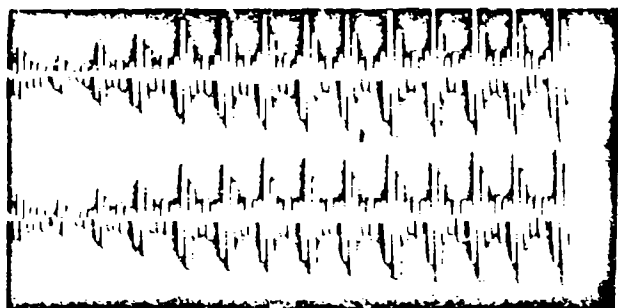
Dotted = Synthetic



a) 119 msec. segment

Synthetic

Original

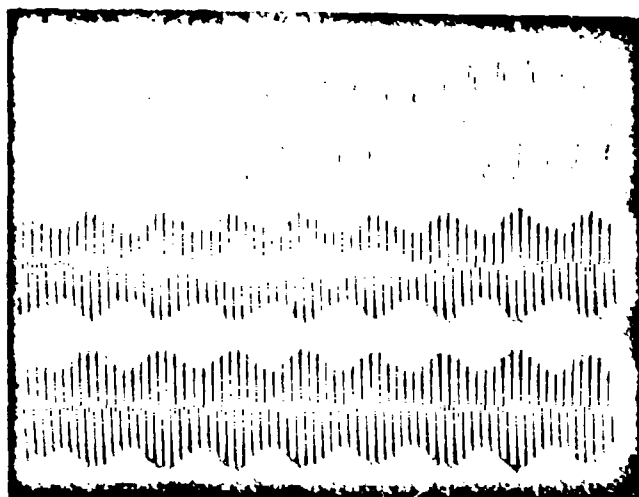


b) 35 msec. segment

Solid = Original
Dotted = Synthetic



Figure C- 19 Synthetic and original waveforms for Band 3 of vowel /a/ as in "all".



b) 21 msec.

a) 70 msec.

Original

Synthetic

Figure C-20 Band 3 synthetic and original waveforms for vowel /a/ as in "all"



Synthetic

Original

Figure C-21 Band 1 of fricative consonant /F/ as in "for"; original and synthetic waveforms.

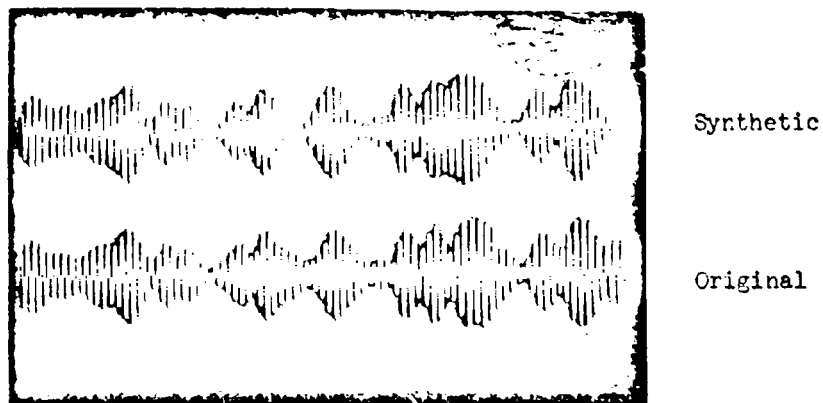


Figure C-22 Band 4 of the original and synthetic waveforms of the fricative consonant /sh/ as in "she".

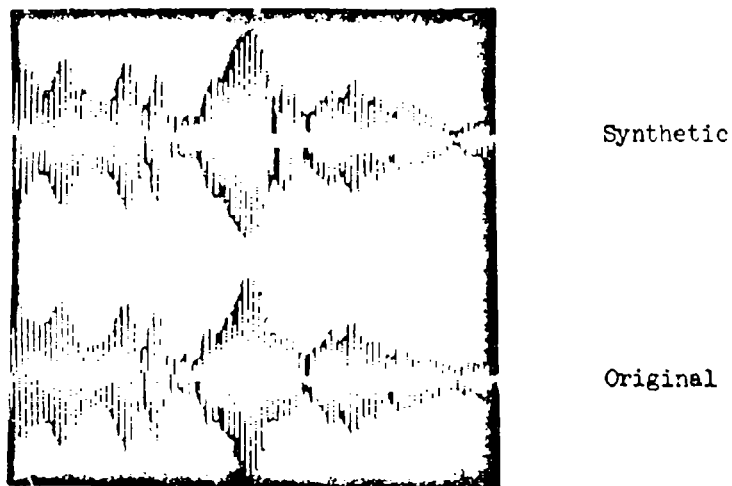
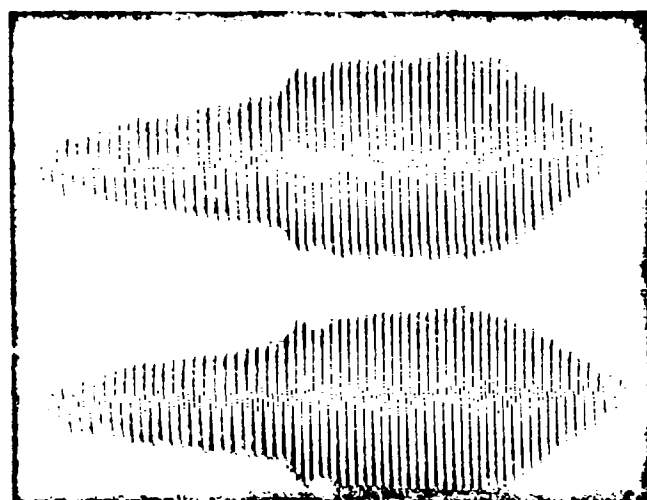


Figure C-23 Band 4 of the original and synthetic waveforms of the stop consonant /t/ as in "to".

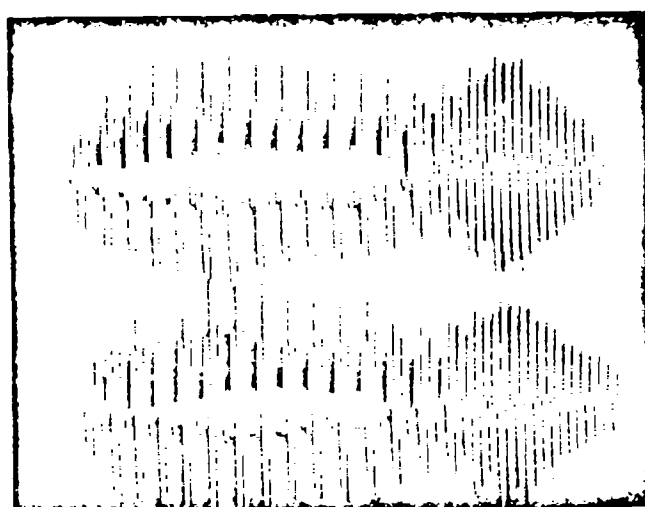
NOT REPRODUCIBLE



Synthetic

Original

Figure C-24 Band 1 of word "me"; female speaker.

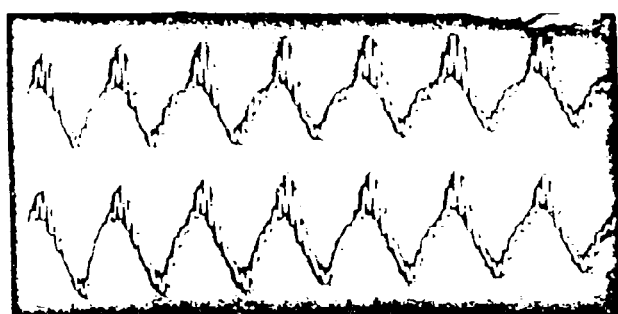


Synthetic

Original

Figure C-25 Band 2 of original and synthetic waveforms of /ei/ portion of word "Pfeifer"; male voice.

Figure C-26 Comparison of 4 Bands and synthetic versus original speech string of 48 word "pete"; female speaker.

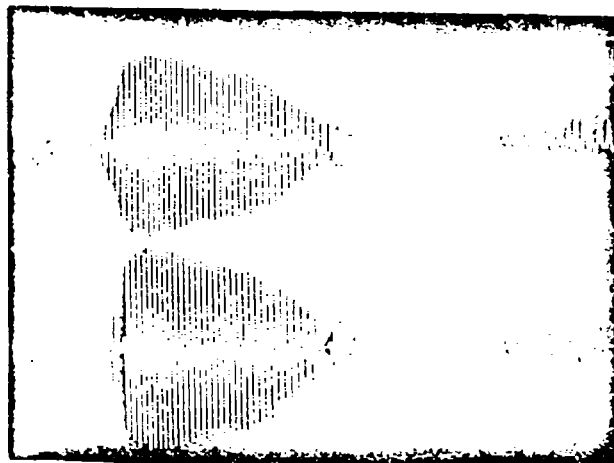


a) Complete original speech string and synthesized speech string; 35 msec.

Original

Synthetic

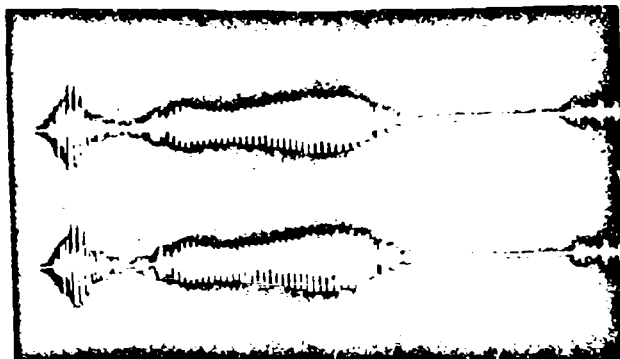
NOT REPRODUCIBLE



b) Band 1; 420 msec. segment

Original

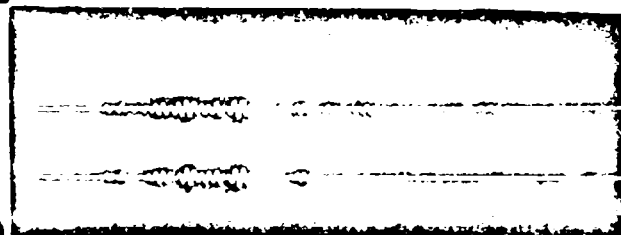
Synthetic



Original

Synthetic

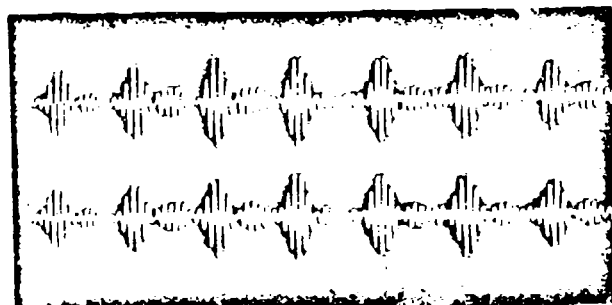
c) Band 2; 420 msec. segment



d) Band 3; 35 msec. segment

Original

Synthetic



Original

Synthetic

e) Band 4; 21 msec. segment

in the SEL-810B speech system currently under implementation.

d) Computer Classification and Recognition of Phonetic Information

The research program in the classification and recognition of phonetic information conducted since the last report has yielded very favorable indications that recognition information can be extracted from the ASCON parameters and used to perform reliable recognition of connected speech. The results which have been obtained which lead to this conclusion are as follows:

1) Extraction of valid frequency information (formant or otherwise) from wave-function parameters.

2) The ability to make a successful vowel map by plotting formant 1 versus formant 2 for steady-state vowels based on frequency data obtained as described in (1) above.

3) Determination of useful fixed-filter bands which can be used for speech recognition.

4) Recognition of steady-state vowels of a single speaker using 3 fixed-filter bands.

5) Recognition of vowels embedded between two unvoiced phonemes for a single speaker, using 3 fixed-filter bands.

6) Preliminary study of the segmentation of connected phonemes using 4 fixed-filter bands.

These results are discussed in greater detail in the following sections.

Extraction of Valid Frequency Information from Wave-Function Parameters

In addition to the normal wave-function parameters, the present analyzer also provides frequency information in the form of a parameter U given by

$$U = \frac{\omega \Delta T}{2}$$

ω = radian frequency

ΔT = 57 microseconds (sampling interval of A-to-D conversion)

From this relation the frequency in Hz. can be calculated as a function of U and represents the frequency of the center of the wave-function

$$f = \frac{U}{\pi \Delta T} \quad (C-18)$$

The average frequency of the wave-function can also be calculated from the wave-function parameters N and S, where

$$f = \frac{N}{S}$$

In working with a wave-function analyzer one of the most important factors in obtaining good parameters is the proper filtering of the raw speech data into appropriate frequency bands or sub-strings. An example of such a sub-string and its corresponding ASCØN and frequency parameters is shown in Figure C-28. It is an 18 ms. portion of the 1200 - 2100 Hz. sub-string of the vowel /æ:/, as in "at". An examination of the frequency data shows that the frequency from one wave-function to the next is not constant, even within one pitch period. This is of course to be expected because of the bandwidth involved and the frequency components which make up the voiced sound.

If it is desired to determine formant frequencies from this type of data then the original speech stream must be filtered into two or three sub-strings, where the bandwidth of each sub-string contains no more than one formant frequency. Assuming the voiced sound is a steady state vowel, then an 18 ms. section like that of Figure C-28 is representative of the entire vowel. A close estimation of the formant frequency within each sub-string can now be

made by taking a weighted average of the frequencies of the wave-functions within the 18 msec. interval. Each frequency term is weighted by the amplitude of the wave-function it represents. Thus, those wave-functions with high amplitudes contribute more power and are given more weight in the frequency equation

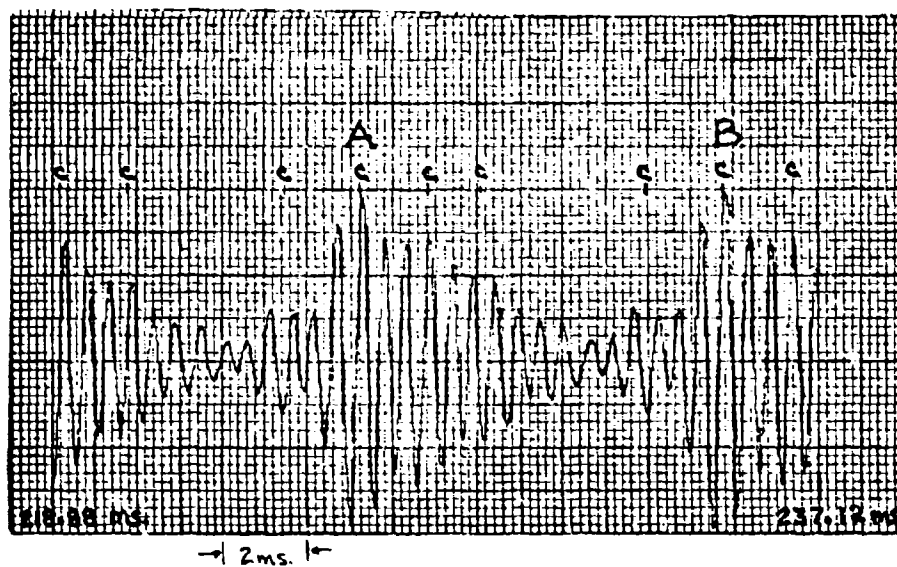
$$F = \frac{A_a f_a + A_b f_b + A_c f_c + \dots + A f}{A_a + A_b + A_c + \dots + A} \quad (C-19)$$

Using the amplitude and frequency data of Figure C-28, formant 2 for that particular vowel was calculated according to Equation (C-18) to be 1835 Hz. Figure (C-29) is a magnitude vs. frequency plot for the same vowel /ae', unfiltered, during the same time interval as Figure C-28. The approximate values of formant 1 and formant 2 are noted on this plot. It should also be noted that the values of formant 2 from the plot and from Equation (C-19) are within approximately 50 Hz. of each other. This kind of close relationship has been demonstrated for the first two formants of all the vowels.

Successful Mapping of Formant Frequencies Derived from Wave-Function

Parameters for a Single Speaker

The key to valid formant frequency calculation from wave-function parameters is the proper filtering around each formant. For connected speech this would imply some sort of automatic formant tracking filter. Since no such tracking filter exists it was necessary to simulate one. The simulation procedure involved taking a fourier transform of a representative sample of a steady-state vowel. Figure C-30a shows an 18 msec. portion of the steady-state vowel /u/, as in "boot". Its corresponding Fourier transform plot is shown in Figure C-30b. From (b) the frequency cutoffs around the first two formants were chosen as 144 - 720 Hz. and 720 - 1700 Hz. Figure C-31 depicts the data



Vowel /ae/ (at), 1200-2100 Hz.

VOWEL "AE" (AT), MID RANGE 1200 - 2100 Hz.
#105

TIME WINDOW = 218.88MS. TO 237.12MS.

A	S(HS.)	C(HS.)	φ(DEG.)	N	F(HZ.)
0.03357	2.44	219.16	-307.9	4.57	1871.2
0.02056	3.25	220.64	-250.2	6.09	1871.2
0.01696	2.16	224.35	-180.0	4.12	1909.4
0.04852	3.43	226.06	-300.6	6.09	1776.5
0.03640	2.10	227.71	-325.9	3.87	1846.5
0.02517	2.73	220.96	-126.6	5.12	1871.2
0.01696	2.22	232.84	-180.0	4.12	1858.8
0.04855	3.45	234.61	-307.3	6.09	1765.2
0.03640	2.63	236.26	-352.7	4.92	1871.2

Figure C-28 18 msec. portion of the vowel /ae/ as in "at" and its corresponding parameters.

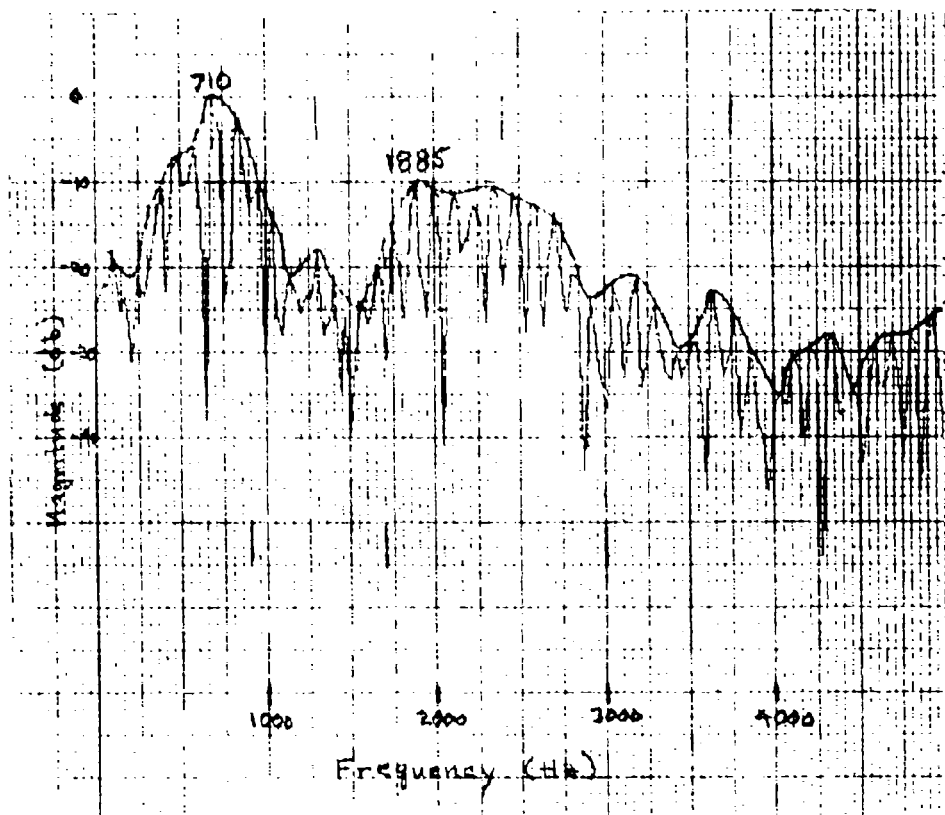
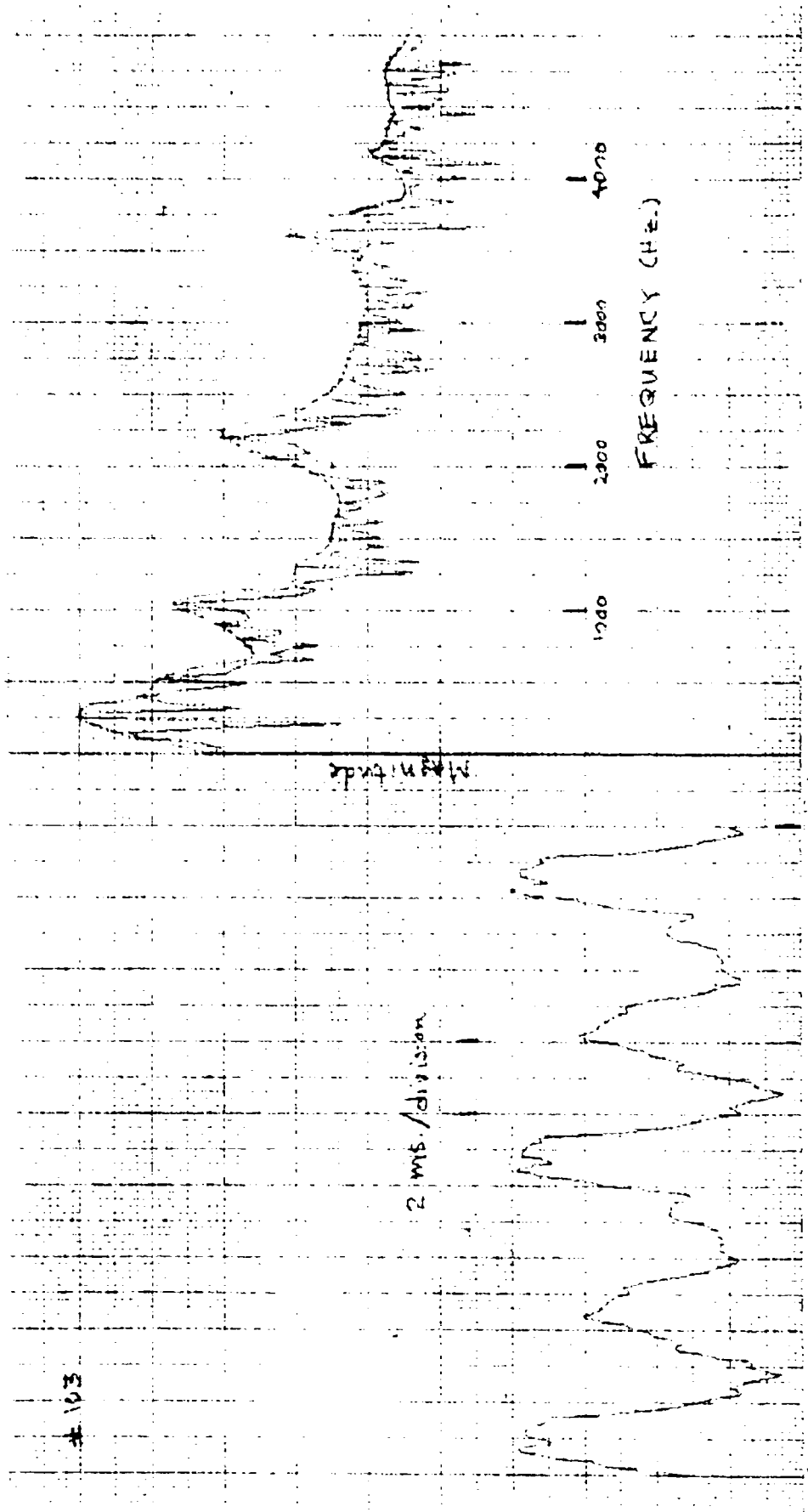


Figure C-29 Magnitude vs. frequency plot for an 18 msec. interval of the steady-state vowel 'ae' as in "at".



Vowel /u/ (boot), 144 - 4000 Hz. Normal Pitch

(a) (b)

Figure C-30 (a) 18 msec. portion of the steady-state vowel 'u', as in "boot" (b) Fourier transform plot of (a)

NOT REPRODUCIBLE

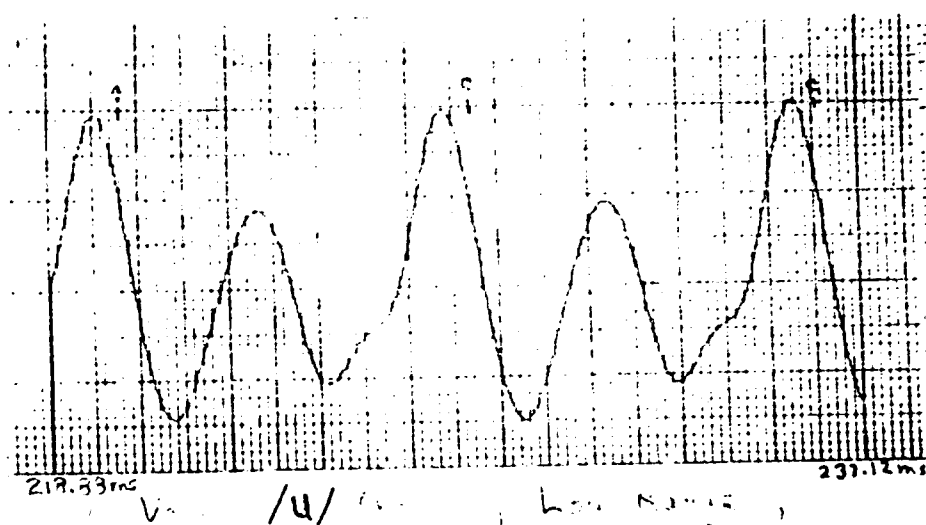


Figure C-31 18 msec. portion of the steady-state vowel /U/ as in "boot" filtered 144 - 720 Hz.

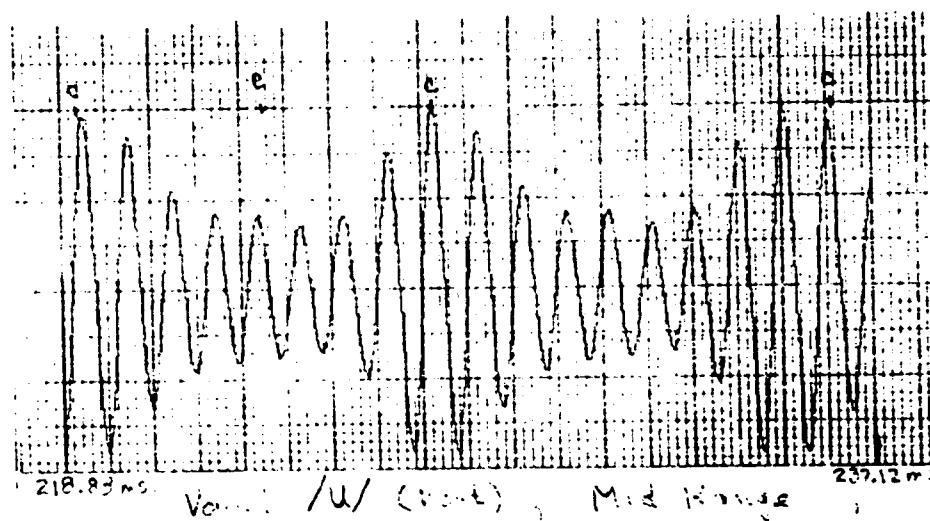


Figure C-32 18 msec. portion of the steady state vowel /U/ as in "boot" filtered 720 - 1700 Hz.

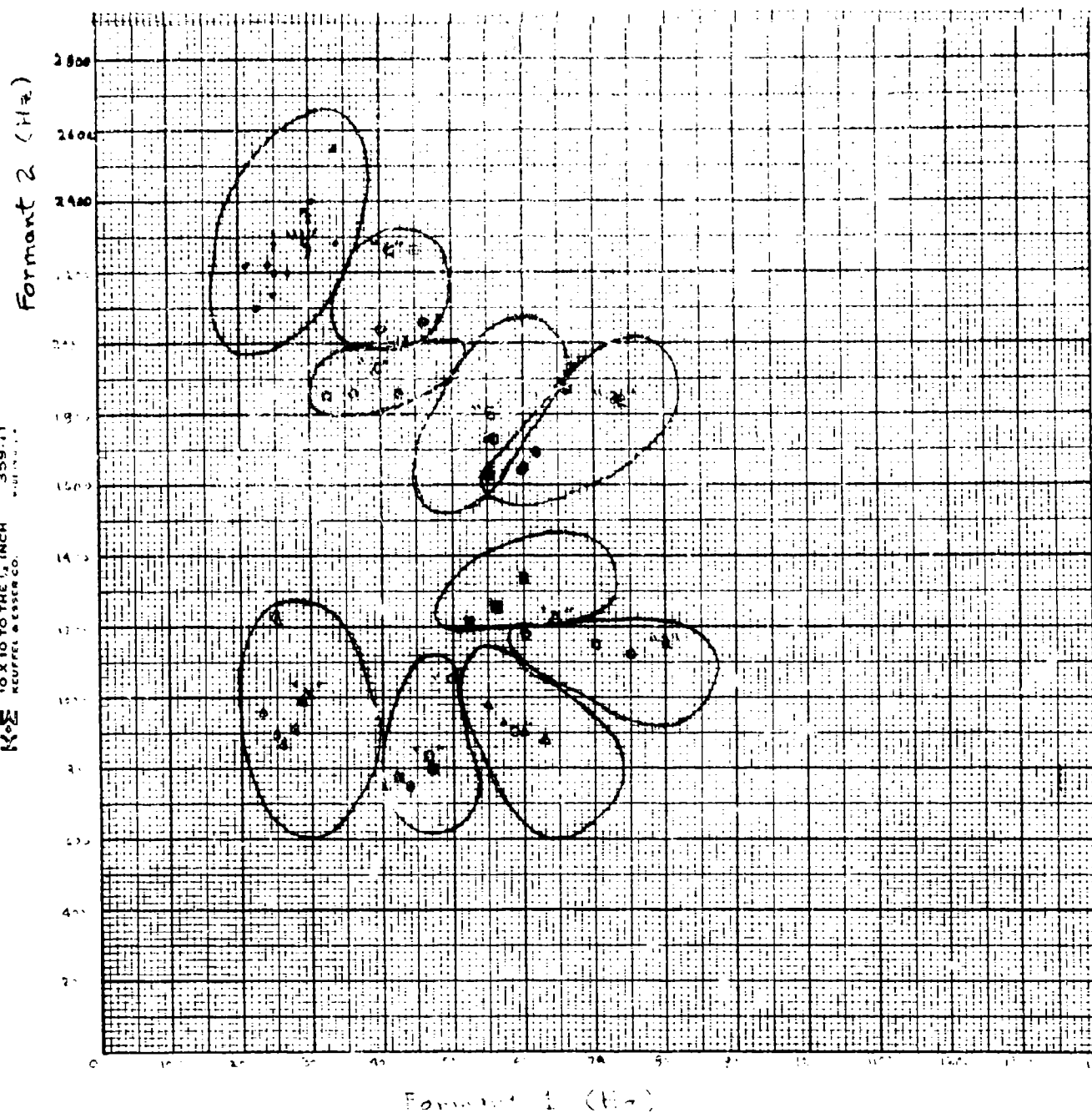


Figure C-33 Formant plot of 10 vowels by a single speaker, where the formant frequencies have been derived from wave-function parameters.

in the 720 - 1700 Hz. band. In both bands the data looks good from the standpoint of having a nice wave-function structure.

The filtering operation was performed using a convolution program in the IBM 1800 computer, resulting in an approximation to an ideal filter. The filtered sub-strings were then analyzed and formant frequencies calculated from the wave-function parameters.

As mentioned the first two formants for 10 steady-state vowels were calculated according to Equation (C-19) and those results were plotted as shown in Figure 33. Each vowel was spoken at least 3 times by a single speaker, and at different pitches. This plot indicates the "vowel loop" for that speaker and shows reasonable separation between the vowels.

Verification of the results obtained from the procedure just described was accomplished by two means. 1) The frequencies calculated from the parameter data were compared with the locations of the formant peaks on the Fourier transform plot. 2) The overall results of Figure C-33 were compared with those of the Peterson and Barney plot (1) which is a plot of formant 1 vs. formant 2 for 76 speakers - men, women, and children.

While the formant map of Figure C-33 looks as though it could provide a good foundation for vowel recognition, there would still remain the problem of filtering. As was mentioned in the discussion of the preprocessor the simulation of an automatic tracking filter required a large amount of computer time in the recognition studies. As a result a set of fixed-frequency parameters were defined and are now being used in the recognition process.

Fixed-Filtering as Related to Speech Recognition

As indicated in the discussion of the preprocessor the fixed filter

parameters which were selected were chosen so that they contained relevant information for recognition. The wave-function parameters resulting from an analysis of the fixed-filter sub-strings for the word "steek" (Figure C-5) are shown in Tables C-1 through C-4. Table 4 is incomplete because of the large number of parameters involved. Besides the five wave-function parameters, each Table contains the average frequency of each wave-function, the difference between the successive principle C parameters (labeled DELTAC), and the difference between successive principle frequency terms (labeled DELTAF).

The principle frequency and principle C just mentioned are those associated with what is called the principle wave-function. The definition of a principle wave-function during a voiced sound is: that wave-function having the maximum amplitude of all the wave-functions within one pitch period. Therefore, there is one principle wave-function associated with each pitch period. In most cases the voiced phonemes in the 100 - 400 Hz. sub-string are made up of one wave-function per pitch period. Therefore every one is a principle wave-function. This can be seen by looking at Table C-1. The principle wave-functions are marked by a star next to the amplitude parameter. DELTAC in this case is a valid pitch period measurement and shows the change in pitch as a function of time, thus providing information on voice inflection and accentuation.

It is also possible to have up to seven or more wave-functions per pitch period. When this occurs, the one with the largest amplitude is classified as a principle, while the remaining wave-functions are classified as followers. Referring back to the sub-string shown in Figure C-28, the principle wave-functions occur at times A and B. The parameter listings of Table C-4 for

RAND 1 OF "STECK" SIN X FILTER KERNEL 100-400 Hz.

24 PARAMETERS
24 SORTED PARAMETERS

A	S(MS.)	C(MS.)	DELTA C	O(DEG)	H	F(HZ.)	DELTA F
0.0506*	14.90	133.03		360.0	2.56	171.4	
0.5973*	7.45	144.21	11.17	140.4	1.75	235.3	63.9
0.6407*	9.15	150.87	6.66	130.8	2.46	269.1	33.7
0.6950*	9.60	157.83	6.95	61.4	2.72	282.0	12.9
0.8062*	8.79	164.27	6.44	49.1	2.46	280.0	-2.0
0.7925*	9.25	170.71	6.44	44.3	2.61	282.5	2.5
0.7797*	9.19	177.09	6.38	43.7	2.61	284.2	1.7
0.7725*	9.21	183.54	6.44	42.1	2.61	283.7	-0.5
0.7540*	9.41	189.98	6.44	41.1	2.66	283.4	-0.3
0.7234*	9.63	196.42	6.44	39.5	2.72	282.9	-0.5
0.7311*	9.66	202.97	6.55	35.0	2.72	282.0	-0.8
0.7129*	9.97	209.50	6.61	33.4	2.78	278.9	-3.0
0.7015*	10.00	216.25	6.66	30.0	2.78	276.2	-2.7
0.6843*	10.61	223.09	6.84	26.1	2.91	274.5	-1.7
0.6821*	10.70	229.99	6.89	22.1	2.91	272.0	-2.5
0.6577*	11.11	236.94	6.95	25.5	2.97	268.0	-3.9
0.6032*	11.17	244.13	7.18	24.1	2.91	269.5	-7.4
0.5465*	12.09	251.54	7.41	22.1	3.05	252.4	-8.1
0.4981*	12.61	259.17	7.63	18.7	3.12	247.6	-4.7
0.4100*	16.66	267.27	8.00	357.2	4.13	247.8	0.1
0.2687*	13.12	276.56	9.20	256.3	3.20	244.2	-3.5
0.1850*	10.82	285.17	8.60	280.8	2.56	236.5	-7.6
0.1046*	12.95	294.17	9.00	322.6	2.78	214.9	-21.5
0.0423*	15.16	384.97	90.80	360.0	3.45	228.2	13.2

Table C-1 Wave-function parameters for the 100 - 400 Hz. sub-string of "steck".

"STEEL" -- 400 - 900 Hz. -- SORTS

34 PARAMETERS
29 SORTED PARAMETERS

NOT REPRODUCIBLE

A	S(HS.)	C(HS.)	DELTA C	O(HS.)	H	F(HZ.)	DELTA F
0.0338*	6.71	132.92		360.0	4.74	706.9	
0.0302*	4.83	132.33	5.41	0.9	2.66	551.3	-155.6
0.1603*	11.99	146.60	8.26	324.3	5.56	463.9	-37.4
0.3079*	10.55	152.70	6.09	17.8	5.33	505.6	41.7
0.4318*	8.24	158.11	5.41	196.8	4.57	554.7	49.0
0.4619*	8.66	164.33	6.21	224.9	4.74	547.0	-7.6
0.4535*	8.05	170.54	6.21	245.4	4.41	548.0	1.9
0.4225*	8.02	176.98	6.44	230.4	4.41	550.3	2.2
0.3723*	7.99	183.42	6.44	227.5	4.41	552.5	2.2
0.3642*	7.47	189.86	6.44	223.7	4.12	552.5	0.9
0.3258*	7.63	196.36	6.49	214.0	4.13	541.7	-10.7
0.2957*	8.17	202.97	6.61	193.1	4.41	540.7	-1.9
0.2820*	7.77	209.58	6.61	180.0	4.13	531.5	-9.2
0.2660*	9.21	216.31	6.72	166.8	4.74	515.0	-16.5
0.2679*	9.86	223.15	6.84	148.1	4.92	499.3	-15.6
0.2653*	8.71	230.16	7.01	129.8	4.41	506.6	7.3
0.2225*	7.47	237.29	7.12	104.0	3.76	503.7	-2.8
0.2006*	8.65	244.75	7.40	60.4	4.26	493.2	-10.5
0.1868*	8.02	252.16	7.41	40.5	3.88	483.8	-9.3
0.1782*	7.95	259.80	7.63	30.8	3.46	490.6	6.8
0.0719	7.61	263.95		258.2	3.12	410.3	
0.1247*	5.87	267.55	7.75	35.0	2.90	404.0	4.2
0.0492	7.45	270.23		357.2	2.90	390.2	
0.0624*	9.79	277.30	9.74	175.4	3.88	396.4	-38.5
0.0443	5.76	282.26		189.9	2.78	482.3	
0.0547*	5.31	286.93	9.63	77.2	2.78	523.5	127.1
0.0537*	5.94	291.66	4.73	85.4	3.12	525.5	2.0
0.0274*	6.03	373.35	81.68	360.0	4.13	684.6	159.9
0.0460*	3.69	380.19	6.84	179.9	2.56	602.0	8.3
0.0532	3.27	382.75		184.5	2.24	686.3	
0.0677*	8.35	387.31	7.12	179.9	4.57	547.0	-145.8
0.0441	8.74	391.97		268.6	4.26	488.1	
0.0366*	4.04	400.31	12.99	234.0	2.41	597.1	50.1
0.0543*	5.04	404.18	3.87	259.1	3.12	619.5	22.3

Table C-2 Wave-function parameters for the 400 - 900 Hz. sub-string of "steek".

NOT REPRODUCIBLE

"STEEL" -- 900 - 1800 Hz. -- SCPTD

20 PARAMETERS
19 SORTED PARAMETERS

A	S(HS.)	C(HS.)	DELTA C	O(DEC)	"	F(HZ.)	DELTA F
0.0360*	1.80	128.64		306.6	2.46	1362.5	
0.0360*	1.00	131.38	2.73	52.6	2.46	1293.5	-69.0
0.0573*	2.26	145.63	14.25	179.0	3.20	1410.4	116.0
0.0317	1.66	147.11		14.0	2.61	1560.1	
0.0573*	2.22	151.06	6.32	179.0	3.20	1459.4	28.9
0.0608*	2.52	158.28	6.32	179.0	3.65	1446.0	7.4
0.0608*	2.58	164.61	6.32	179.0	3.65	1417.6	-29.3
0.0602*	2.52	170.82	6.21	233.3	3.55	1410.4	-7.1
0.0485*	2.65	177.32	6.49	179.0	3.76	1417.6	7.1
0.0477*	2.21	183.82	6.49	127.3	3.12	1410.4	-7.1
0.0337*	3.36	190.15	6.32	179.0	4.74	1410.4	0.0
0.0337*	3.41	196.50	6.44	179.0	4.74	1389.4	-20.0
0.0365*	2.13	203.09	6.49	179.0	2.97	1396.4	6.0
0.0365*	2.13	209.70	6.61	179.0	2.97	1396.4	0.0
0.0332*	2.33	216.20	6.49	269.0	3.20	1309.2	-27.2
0.0332*	2.40	222.98	6.70	269.9	3.20	1330.1	-39.0
0.0355*	2.19	229.09	7.01	179.9	2.97	1355.9	25.7
0.0332*	2.46	236.89	6.89	200.0	3.20	1299.5	-56.4
0.0306*	1.99	244.24	7.35	179.0	2.56	1342.0	43.4
0.0331*	2.42	302.81	130.56	90.0	3.28	1355.9	12.0

Table C-3 Wave-function parameters for the 900 - 1800 Hz. sub-string of "steel".

"STEEL" -- 1800 - 3600 "Z. -- SORT9

62.

201 PARAMETERS
75 SORTED PARAMETERS

	A	S(MS.)	C(MS.)	DELTA C	O(DEC)	"	F("Z.)	DELTA F
(a)	0.0274*	1.26	20.40		179.9	4.12	3263.8	
	0.0332*	1.00	21.31	0.91	89.0	3.20	3189.8	-73.9
	0.0338*	1.48	33.34	12.02	360.0	4.74	3189.7	-0.1
	0.0338*	1.45	34.82	1.48	360.0	4.74	3264.0	74.3
	0.0455*	1.22	35.91	1.08	86.8	3.87	3153.9	-110.0
	0.0352	1.11	36.76		181.8	3.76	3321.7	
	0.0297*	0.92	37.27	1.36	216.6	3.12	3301.9	228.0
	0.0344*	1.28	38.07	0.79	3.6	4.12	3226.3	-155.6
	0.0274	1.33	44.57		179.9	4.12	3084.5	
	0.0338*	1.53	45.88	7.80	360.0	4.74	3084.5	-141.7
	0.0332*	1.00	56.25	10.37	269.9	3.20	3189.7	105.1
	0.0274*	1.25	58.25	1.99	179.9	4.12	3302.3	112.6
	0.6226*	1.45	171.85	6.27	360.0	3.87	2673.3	121.5
	0.3604	0.68	172.65		78.6	1.85	2699.0	
(b)	0.3706	1.14	173.39		338.8	2.97	2599.0	
	0.2325	0.55	174.97		168.6	1.50	2699.0	
	0.2365	1.00	174.76		34.0	2.72	2699.0	
	0.0342	0.73	175.44		211.8	2.28	3118.9	
	0.1235	1.13	176.07		224.9	3.28	2893.6	
	0.0033	1.20	176.87		269.9	3.12	2599.0	
	0.6318*	1.40	178.23	6.38	335.8	3.76	2673.3	0.0
	0.4370	0.92	179.03		75.9	2.16	2339.1	
	0.3797	1.09	179.83		201.1	2.61	2378.8	
	0.1909	2.54	181.20		269.9	6.09	2399.1	
	0.2277	1.41	181.88		122.1	3.55	2506.1	
	0.2197	1.08	182.68		226.2	2.56	2358.7	
	0.1856	0.89	183.36		32.4	2.28	2551.8	
	0.5317*	1.65	184.67	6.44	321.0	4.41	2673.3	0.0
	0.3784	1.13	185.42		41.1	2.72	2399.1	
	0.2334	2.71	186.33		89.0	5.82	2142.7	
	0.2005	1.98	187.64		195.0	2.72	2506.1	
	0.2374	3.59	188.44		288.7	8.00	2227.8	
	0.1090	1.28	189.24		25.1	2.97	2319.6	
	0.5246*	1.97	191.06	6.38	350.0	5.12	2599.0	-74.3
	0.3396	1.09	191.86		25.5	2.56	2339.1	
	0.2597	1.04	192.66		184.5	2.46	2358.7	
	0.2195	1.03	193.34		16.8	2.46	2378.8	
	0.1863	1.55	194.14		134.9	3.76	2419.7	
	0.2074	2.17	195.62		88.1	5.12	2358.7	
	0.1450	1.36	196.59		1.8	3.36	2462.1	
	0.5151*	1.92	197.56	6.49	308.9	4.74	2462.1	-136.8
	0.3019	1.31	198.36		360.0	2.97	2263.6	
	0.2351	3.14	199.44		266.3	7.11	2263.7	
	0.1938	1.52	200.81		330.1	3.45	2263.6	

Table C-4 Wavefunction parameter for the 1800 - 3600 Hz. sub-string of the word "steek".

- (a) a portion of the /s/
(b) a portion of the /ee/

the vowel portion of the word "steek" likewise show the relationship between the principle wave-functions and the followers.

During an unvoiced phoneme and some phoneme transitions the wave-functions are somewhat random or burst-like. For these cases there is no strict principle-follower definition. Therefore, a wave-function is selected as a principle if its amplitude is within 20 % of the amplitudes of the wave-functions both before and after it or if its amplitude is greater than 80% of the amplitude of the previous principle wave-function.

For the word "steek", Band 1 is described by 24 parameter sets, Band 2 - 34 parameter sets, Band 3 - 20 parameter sets, and Band 4 - 201 parameter sets. The total for the entire word is 279 parameter sets. If the principle wave-functions are sorted from each band then Band 2 is reduced to 29 parameter sets, Band 3 - 19 parameter sets, and Band 4 - 75 parameter sets. This results in a total of 147 parameter sets for the whole word. The usefulness of these sorted parameters will be shown later.

Recognition of Steady-State Vowels of a Single Speaker Using Three Fixed-Filter Bands

The algorithm for the recognition of steady-state vowels was implemented in FORTRAN on the IBM 1800 computer. Due to limitations in memory, however, only three bands could be operated upon, but the results were still quite successful. A frequency parameter was calculated for each sub-string using Equation (C-10) and wave-function parameters for an 18 msec. portion of the vowel. This was not a true formant frequency but it was used in the same manner. An amplitude parameter was assigned to each substring by extracting the amplitude parameter for the first principle wave-function occurring within the same 18 msec. interval. Examination of the frequency and amplitude data for 12 vowels revealed

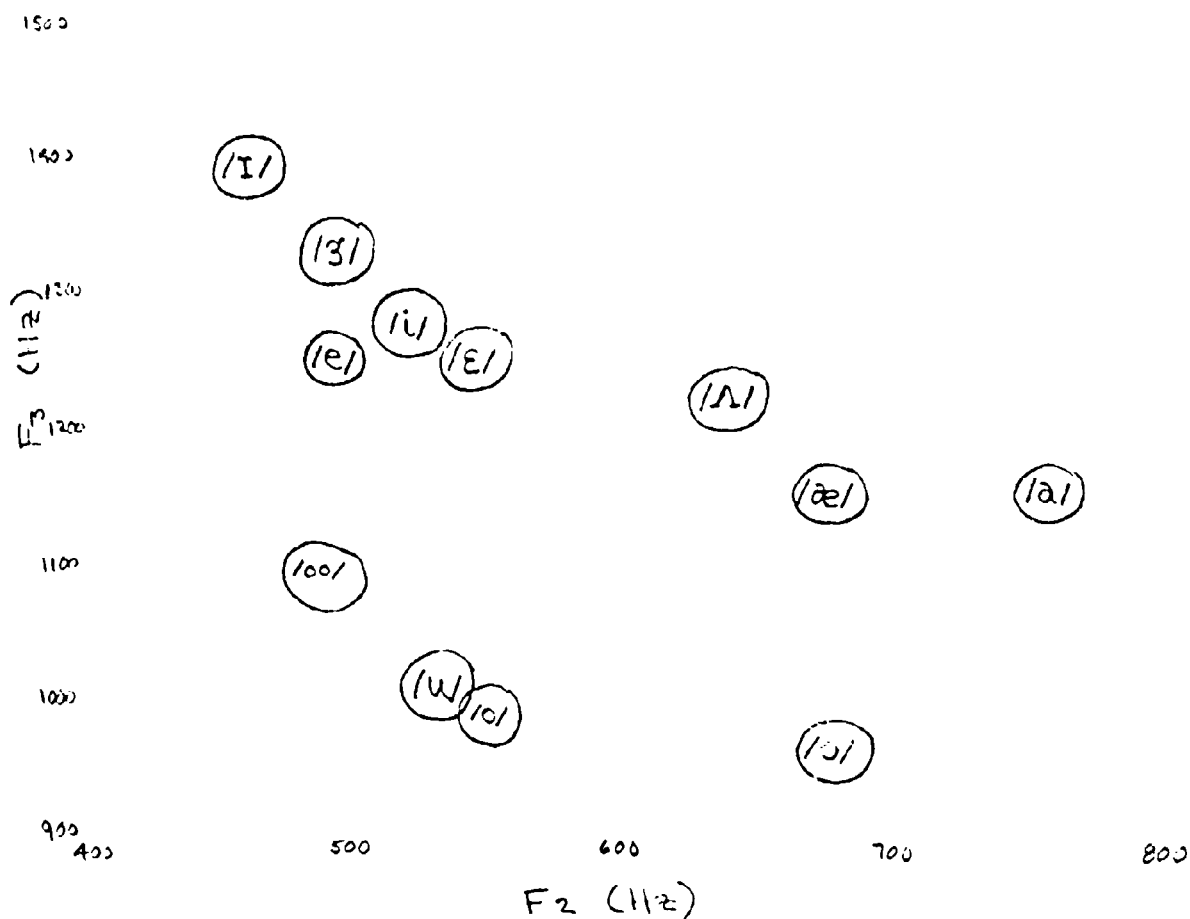


Figure C-3⁴ Steady-state vowel frequency plot.

F_2 = Weighted average frequency in 400 - 900 Hz. Band

F_3 = Weighted average frequency in 900 - 1800 Hz. Band

Single Speaker

that the frequency in Band 1 did not vary significantly to warrant its use as a recognition parameter. Because of the higher formant content of Bands 2 and 3 the frequency parameters for these bands were plotted against each other and yielded a useful map with reasonable separation in most cases.

(See Figure C-34)

A second vowel map was made utilizing the amplitude parameters. In most cases the amplitude relationships between sub-strings seemed to vary from vowel to vowel. When the amplitudes in Bands 2 and 3 were first normalized by the amplitude of Band 1 and then plotted against each other, the result was the second map, with reasonable separation between vowels. (See Figure C-35).

The recognition algorithm itself was based on a simple binary decision tree. Decision lines were drawn through the frequency map, isolating each of the vowels or groups of vowels as much as possible. The final recognition decision was then made after an examination of the amplitude map, on which decision lines were also made. The decision tree used is shown in Figure C-36.

For a single speaker the results were more than 95% correct on the 12 vowels making use of only the first three fixed filter bands. An exhaustive study and test was not conducted because of the limited usefulness of a steady-state vowel recognizer. However, each vowel was spoken from 2 to 3 different ways, and the vowel /i/, as in "beet", was spoken 12 times at various pitches. Enough information was gathered and the results conclusive enough to demonstrate the feasibility of using wave-function parameters for phoneme recognition. The next step seemed to be a more realistic one, that of vowel phonemes connected to other phonemes.

Recognition of Vowel Phonemes Embedded between Two Unvoiced Phonemes for a Single Speaker Using 3 Fixed-Filter Bands

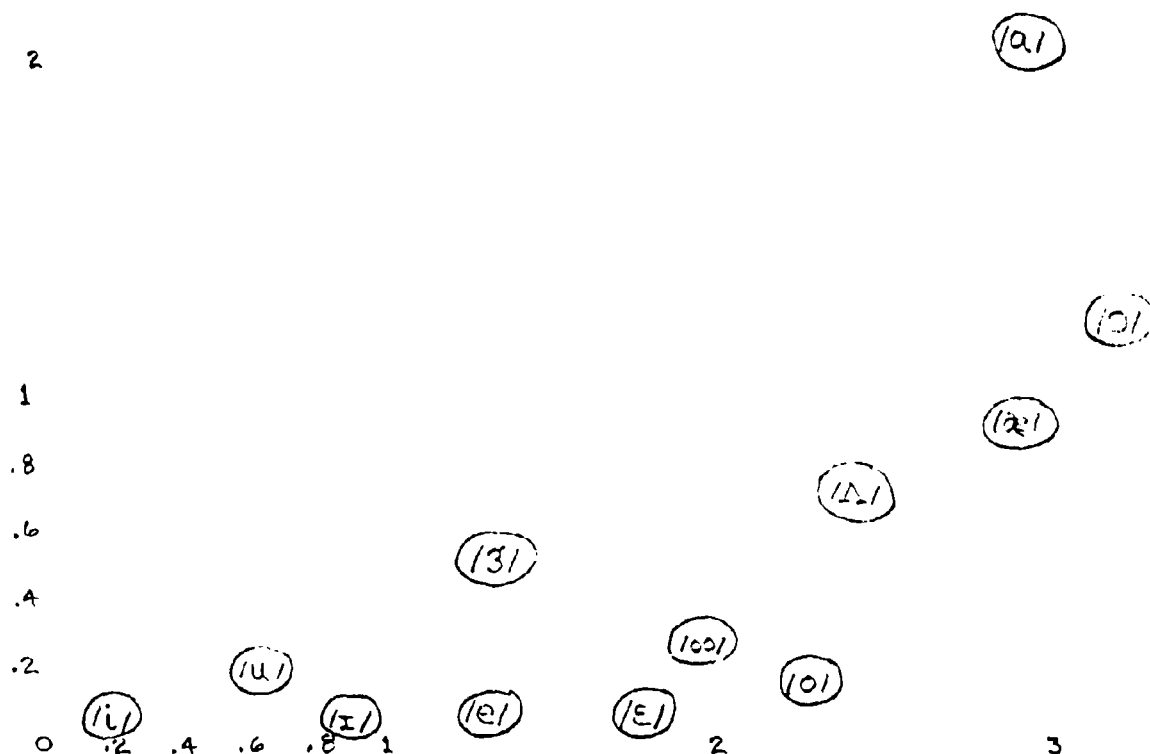


Figure C-35 Steady-state vowel map as a function of Band 2 and Band 3 amplitudes.

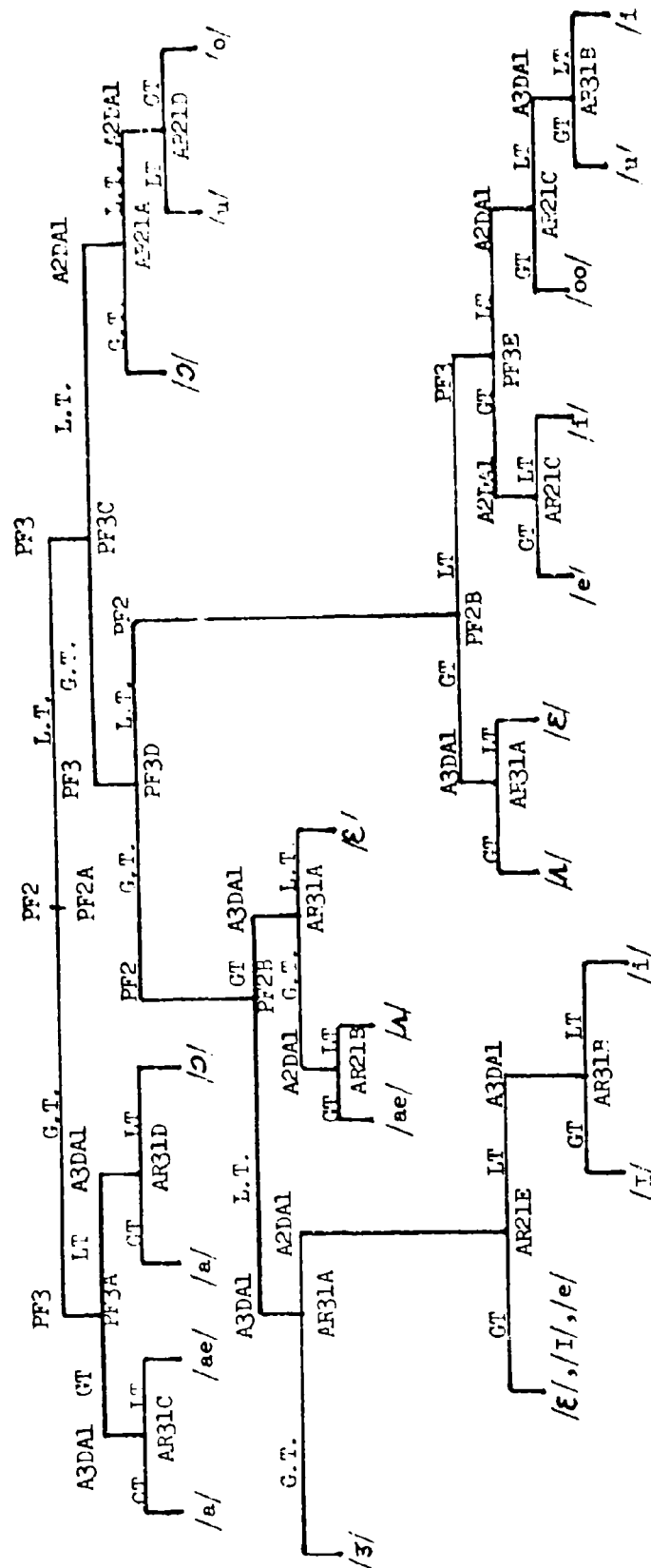


Figure C-36 Decision tree for steady-state vowel recognizer utilizing 3 fixed filter bands. Decisions are made on 4 parameters: (1) Weighted average frequency in Band 2, (2) Weighted average in Band 3 (3) Amplitude of Band 2 normalized by amplitude of Band 1 (4) Amplitude of Band 3 normalized by amplitude of Band 1.

The idea of embedded vowel recognition suddenly becomes complicated by the following facts:

- 1) With steady-state vowels the formant frequencies remained fairly constant throughout the duration of the vowel. Therefore, the weighted average frequency associated with each filter band remained fairly constant. Now the formants (and therefore the average frequency within each band) vary as a function of time, depending upon the phonemes before and after the vowel and the amount of coupling between them.
- 2) The amplitude of steady-state vowels remained fairly constant throughout the vowel. Now the amplitude is generally some form of increasing-decreasing function.
- 3) Because of the continuous nature of the steady-state vowel, almost any time-interval could be used as a representative sample of that vowel. Now the time during which the vowel occurs must first be determined before recognition information can be derived from the wave-function parameters.

Due to the use of FORTRAN, this recognition scheme was also restricted to the first three fixed-filter bands. Because of the good results of the steady-state vowel recognizer, it was decided to use amplitude and frequency as the recognizer inputs. The only restriction on the spoken word was that the phoneme content be unvoiced-voiced vowel-unvoiced.

The first step in the process was to do a sort on the first three bands. The sort program, as described previously, finds the principle wave-function associated with each pitch period and discards the rest. This indicates a pitch synchronous type of recognition. A scan was then made of Band 1 to determine the time when the pitch phenomenon began and ended, the result

being the time during which the vowel occurred.

This time window was then projected to Bands 2 and 3. If there were more than 5 pitch periods during the vowel, the first two and last two wave-functions were discarded in order to help eliminate phoneme transition effects. A simple average was then calculated for the remaining principle frequencies in Bands 2 and 3.

It has been noted that the principle wave-functions in each band are within alignment of each other by ± 3 msec. To perform valid amplitude normalization, the amplitude of each principle wave-function occurring during the vowel in Band 2 was divided by the amplitude of the corresponding principle wave-function in Band 1. A simple average was then taken of the resulting amplitude ratios, yielding an overall A_2/A_1 . The same procedure was followed to determine A_3/A_1 .

An example of the procedure is as follows. Figure C-37 is a picture of the vowel portion of the word "sock" and its first three corresponding sub-strings. After an analysis and sort on each band, a scan was made of Band 1 in order to determine the time occurrence of the vowel segment. The vowel time-window is shown at the top of Figure C-38. Also in the figure are the principle A, C, and F parameters for the first three bands, which occurred during the defined time window. An alignment procedure was then performed. Each wave-function in Band 1 was compared, in time, with those of the other two bands. Those that occurred within 3 msec. of each other in all three bands were saved, the others discarded. The A, C, and F parameters remaining after such an alignment are shown in Figure C-39. From these data, each amplitude parameter in Band 2 was divided by the corresponding amplitude

parameter in Band 1, yielding a set of A_2/A_1 ratios. The same was done to generate a set of A_3/A_1 ratios, and these two ratio sets are shown in Figure C-40. Also given in that figure are the simple averages of these two sets, thus giving two recognition parameters. Figure C-4 also shows the results of taking the simple average of the frequency terms of each band contained in Figure C-39. The average frequencies calculated for Bands 2 and 3 were also used as recognition parameters.

As with the steady-state vowel recognition the two amplitude terms were plotted against one another and the two frequency terms resulting in two embedded vowel maps. Decision lines were drawn first on the frequency map to isolate the vowels as much as possible and the final recognition decisions were based on the amplitude map. A binary decision tree was then used as the basic recognition algorithm, based on the decision information extracted from the two vowel maps.

The results of this recognition technique were good though not as impressive as steady-state vowels. Out of 32 words, each containing one vowel, only 4 wrong decisions were made. Words were chosen such that each of 12 vowels was done at least twice. Some of the words spoken were: coast, sit, hat, sock, foot, etc.

Misrecognition occurred on the vowels embedded in the following words: "gus", "get", "tug", and "shook". The reason for errors on the first three can be attributed to the fact that the vowel in each was coupled with a voiced /g/ phoneme. Therefore, transition effects were included as part of the vowel and resulted in erroneous recognition parameters.

Because of errors like this it became clear that the first step in

NOT REPRODUCIBLE

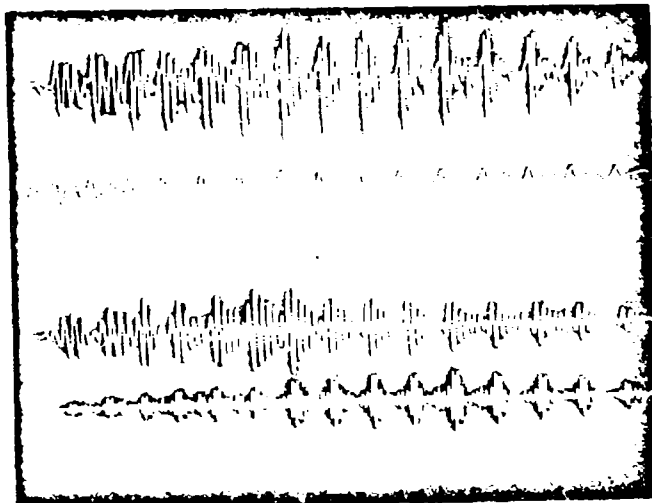


Figure C-37

Vowel portion of the word "sock" and its first three sub-strings.

179.8350 266.9880

A	C(msec.)	F(Hz.)
0.2843	179.835	357.5
0.2653	188.157	342.7
0.2563	196.421	347.6
0.2455	204.800	349.5
0.2097	213.179	334.0
0.2025	221.573	331.6
0.2058	230.280	284.9
0.1900	239.001	293.3
0.1545	248.178	279.8
0.1249	257.355	290.7
0.1045	266.988	274.8

B1

100 - 400 Hz.

A	C(msec.)	F(Hz.)
0.8173	189.183	701.6
0.7946	197.448	714.0
0.7758	205.656	719.7
0.6473	214.035	734.7
0.5872	222.528	715.9
0.5628	231.249	708.3
0.4581	240.140	698.2
0.3061	249.204	669.7
0.2140	258.609	669.7

B2

400 - 900 Hz.

A	C(msec.)	F(Hz.)
0.6496	189.126	1150.3
0.6721	197.390	1194.3
0.6491	205.713	1179.3
0.7686	214.148	1179.3
0.6035	222.585	1162.4
0.5535	231.363	1189.4
0.4610	240.198	1194.3
0.2883	249.318	1179.3
0.2409	258.609	1220.4

B3

900 - 1800 Hz.

Figure C-38 Principle A, C, and F parameters occurring during the vowel portion of the word "sock".

BAND 1			BAND 2			BAND 3		
A	C(msec.)	F(Hz.)	A	C(msec.)	F(Hz.)	A	C(msec.)	F(Hz.)
0.256	224.238	347.671	0.587	222.528	715.965	0.603	222.585	1169.471
0.245	232.617	349.887	0.562	231.249	708.807	0.553	231.363	1189.411
0.209	240.996	334.037	0.458	240.140	698.240	0.461	240.198	1194.354
0.202	249.489	331.651	0.306	249.204	671.483	0.288	249.318	1179.356
0.203	258.096	285.465	0.214	258.609	660.779	0.240	258.609	1220.429

Figure C-39 Principle A, C and F parameters which are within 13 msec. of each other in all three bands, for the vowel portion of the word "sock".

A_2/A_1	A_3/A_1	F_1 (average)	F_2 (average)	F_3 (average)
2.2903	2.3538	329.7426	692.8548	1190.6048
2.2923	2.2545			
2.1840	2.1985			
1.5149	1.4269			
1.0503	1.1820			
Avg. A_2/A_1	Avg. A_3/A_1	(b)		
1.8663	1.8832			
(a)				

Figure C-40 (a) Amplitude ratios and their averages as calculated from the amplitude parameters of Figure C-39.
(b) Average frequencies of each band as shown in Figure C-39.

this procedure should be expanded before any more recognition was attempted. This step was the one which determines the time during which the vowel occurs. It was decided to make this a general purpose segmentation, where the time occurrence of each phoneme would be defined.

The development of a general purpose segmentation algorithm is currently underway. It will be implemented on the IBM 1800 speech system using Assembly language thus eliminating some of the memory restrictions incurred with FORTRAN on this computer. As a result the data from all four fixed-filter bands will be utilized for the segmentation and later for recognition.

Because of the high information content of these four frequency bands, as described earlier, a highly accurate phoneme segmentation will be possible. This means that phoneme transition effects can be reduced or eliminated. Recognition can then be started with the vowel set and easily be expanded to the entire phoneme set.

(e) Data Compression Studies

The study of data compression is closely related to the problem of speech recognition, with the possibility that some techniques (e.g. segmentation) may be shared. Therefore, for both compatibility and convenience, the frequency bands selected for the ASCON streams were chosen to be those used for the recognition studies. The basic speech waveform is thus assumed to have been bandlimited to .1- 3.6 kHz., and the four sub-bands employed are .1- .4, .4 - .9, .9 - 1.7, and 1.7 - 3.6 kHz.

Two general areas of study have been emphasized in the preliminary compression studies completed to date. The first is the tabulation of the data rate of the "highest fidelity" ASCON representation. This involves making

the best possible fit to the speech waveform, in the time domain, with Gaussian wave-functions and determining the required bit rate without further compression. This is estimated by computing the number of wave-functions required and assuming eight bits for each of the five parameters. Studies of vowels have shown bit rates on the order of 50 - 60,000 bits/second for the raw ASCON stream. This improves somewhat for full words (approx. 30 - 50,000), and even more for phrases, since the ASCON representation automatically ceases for any quiet periods, no matter how short.

The second general area of study has been an attempt to determine how many of these wave-functions are superfluous from a perceptual standpoint. Since the elimination of redundant parameter sets is extremely simple to implement (given a criterion for redundancy), this seemed a natural place to begin rather than to initially attempt those approaches involving encodement of the ASCON sets. A sort program has been implemented to select the "principal"* wave-functions in the four frequency bands, and to reconstruct a replica of the sound based only on the principal wave-functions. In the lower two bands, most of the wave-functions qualify as principals; however, there is considerable redundancy in the two higher bands. Table C-5 illustrates results for a typical vowel sound /i/ and for the word "dirt". The sounds of the reconstructed waveforms for the two cases are perceptually the same whether all the ASCON sets are employed or only the principals. It should be noted that no reduction at all was made in the first two bands. This result can be improved on by the

* The appropriate definition of a principal wave-function is still a subject of study. Currently it appears that if the amplitude of a wave-function is either a local maximum or within 80% of one of its neighbors, it should be so classed.

/1/

Frequency Bands	Number of ASCON Sets	
	Original	After Sort Program
100 - 400 Hz.	47	47
400 - 900 Hz.	65	65
900 - 1700 Hz.	137	42
1700 - 3300 Hz.	320	100
Total	569	254
Data Rate bits/sec.	54,500	24,200

"Dirt"

Frequency Band	Number of ASCON Sets	
	Original	After Sort Program
100 - 400 Hz.	46	46
400 - 900 Hz.	85	85
900 - 1700 Hz.	116	35
1700 - 3300 Hz.	176	54
Total	423	220
Data Rate bits/sec	40,400	21,000

Table C-5

proper definition of a principal wave-function. However, the improvement to the total bit rate is small, and was ignored for this illustration.

In addition to the sort on principal wave-functions, experiments are being conducted to evaluate the discarding of ASCON sets with an amplitude less than a (normalized) threshold. The threshold has been normalized in one of several ways. Since the largest amplitude is generally normalized, the simplest procedure is to set the threshold to a fixed value. This implies that a wave-function will be discarded if the amplitude is less than some specified fraction of the largest wave-function occurring in any of the four bands. A somewhat more effective procedure also normalizes with respect to the largest amplitude in the band and sets a threshold test (at perhaps a different level). This allows a consideration of the possible differences in amplitudes from band to band. These procedures appear to be capable of reducing the data rate by an additional factor of two without changing the perceptual content of the sound. A final procedure, somewhat more cumbersome to implement, is to normalize with respect to the largest amplitude in a moving time window and employ a threshold test. This has so far resulted in a disappointingly small improvement over the previous cases.

As soon as a reliable segmentation program is developed, work will commence on encoding ASCON sets during vowel-like sounds, perhaps by n-level delta modulation techniques, using the segmentation as the guide as to when to start and stop the encoding procedure.

(f) Interrelationships between a Wave-Function Representation and a
Formant Model of Speech^{*}

Theoretical and empirical investigations have been conducted which interrelate

* Markel, John, "On the Interrelationships between a Wave-Function Representation and a Formant Model of Speech", PhD Dissertation, Univ. of Calif., Santa Barbara, July, 1970.

the parameters of the wave-function representation to those of a classical formant model. Two points should be made regarding these relationships. First, the transformations are one way; that is, parameters of the wave-function model are transformed to the formant model only. The reason is that the formant model defines what can be considered as a fundamental set of parameters for voiced speech. In terms of information theory, there is no other known set of acoustic parameters which is capable of describing the essential character of the vowel sounds with lower information capacity. It has been shown that, in general, description of a vowel sound in terms of the wave-function parameters requires many more parameters. What has been developed, then, is several many-to-one transformations which map the wave-function parameter set into estimates of the formant parameter set.

The second point to be made is that the interrelationships to be presented are empirical ones based upon reasonable engineering assumptions along with some theoretical justification. As will be shown later, these relationships have given very good results in predicting parameters of this formant model.

In order that the true parameters might be known, the study was conducted on a set of synthetic vowels. (The procedures have also been applied to real speech, and give reasonable results, but since all procedures for estimating the formant parameters are subject to error, analyzing real speech cannot give a base of reference for error analysis.) The synthetic speech which was analyzed had a realistic glottal driving function, periodicity, a radiation and a correction term.

A final comment relates to preprocessing of the vowels before analysis.

It has been observed that a sufficient condition for wave-function isolation is that major energy regions be isolated during the analysis. Although the regions were separated manually for this study, it appears reasonable to assume that this separation could be accomplished automatically except for the back vowels such as /a/ and /ɔ/. Suzuki* has considered one approach to automatic separation of formant regions by moment methods.

For the cases where two closely spaced formants cannot be resolved automatically, a separate algorithm was developed for estimating both formants and bandwidths from the wave-function parameters that define the region containing two formants. As for the filters, three contiguous $\sin x/x$ type of filters are used to define the range (0,3000) Hz. for each vowel (except for /i/ which has a range of (0,3500) Hz.).

The proposed method for estimating formant parameters from wave-function parameters depends upon each formant region being isolated (except where two closely spaced formants are known to reside within a single filtered region). Also the success of the method depends upon being able to isolate single pitch periods. This requirement is necessary for the estimation of bandwidth. If only estimation of the formant frequencies was desired this requirement could be eliminated.

The first step in estimating formant parameters from wave-function parameters is to isolate a single representative pitch period of the synthetic vowel generated from the wave-function parameters. The set of parameters representing the vowel is then used as the input set for the transformation equations. The following algorithm is used for isolating a single pitch period from the wave-function parameter sets.

* Suzuki, J., Y. Kadokawa, and K. Nakata, "Formant-Frequency Extraction by the Method of Moment Calculations," JASA, Vol. 35, September 1963, pp. 1345-1353.

Over a time interval containing at least one pitch period, the A parameter list is searched for a maximum. (The number of parameter sets considered is determined by noting that the corresponding C parameters must be within the chosen time interval). The five parameter sets corresponding to this maximum A is defined by

$$\Omega(1,n) = [A(1,n), S(1,n), C(1,n), \phi(1,n), F(1,n)]$$

where n denotes the particular filtered region $n = 1, 2, 3$. The start of the next period is determined by finding the maximum A parameter whose corresponding C parameter satisfies $T_{\text{MIN}} < C - C(1,n) < T_{\text{MAX}}$ where T_{MIN} and T_{MAX} define minimum and maximum expected pitch periods. Consider this set $\Omega(M+1, n)$. For region n, the parameter set defining the isolated pitch period is then given by

$$\Omega_n = [\Omega(1,n), \dots, \Omega(M_n,n)]$$

where M_n in general is different for each n. Note that in general, the pitch period will be slightly different for each region.

One other fact that needs to be emphasized is that this algorithm is not being proposed as a useful method for extraction of fundamental frequencies from arbitrary speech. It is simply a reasonable method for extracting the pitch period from the Digital Vowel Synthesizer (DVS) automatically, where all glottal pulses are defined as identical. This is certainly not the case in real voiced speech. The problem of estimating formant parameters can be formulated in the following way. For region n, $n = 1, 2, 3$ the wave-function representation of the isolated vowel pitch period will be

$$\hat{s}_n(t) = \sum_{i=1}^{M_n} \psi[C(i,K)t] \quad (C-20)$$

where

$$\psi[\Omega(i,n),t] = A(i,n) \exp \left[-\pi^2 (t-C(i,n))^2 / S^2(i,n) \right] \cdot \\ \cos[2\pi F(i,n)t - \phi(i,n)]$$

The fourier transform of these wave-function components is

$$\psi[\Omega(i,n),jf] = \frac{A(i,n)S(i,n)}{2\sqrt{\pi}} \exp \left\{ [f - F(i,n)]^2 S^2(i,n) - \right. \\ \left. j[\phi(i,n) + 2\pi f C(i,n)] \right\}$$

From this last expression, note that near $f = F$, the maximum value of the magnitude spectrum is obtained as approximately

$$|\psi(\Omega, jF)| = \frac{AS}{2\sqrt{\pi}} \quad (C-21)$$

The equation says that the peak value of the spectrum of any wave function is proportional to its AS product. This leads to the conjecture that over any time interval P, the importance of any wave-function in relation to the overall spectral result can be ranked in terms of the individual AS products of each wave-function. With this information, the conjecture can be made that each of the F terms contributes to the formant frequency roughly in proportion to its corresponding AS product giving the formant frequency estimator

$$\hat{F}_n = \frac{\sum_{i=1}^{M_n} A(i,n)S(i,n)F(i,n)}{\sum_{i=1}^{M_n} A(i,n)S(i,n)}, \quad n = 1,2,3 \quad (C-22)$$

The estimation of formant bandwidth has traditionally been a frequency domain operation with few attempts in the time domain. A problem which exists in frequency domain methods is that of constantly overestimating the bandwidth value. Even if the discrete Fourier transform is obtained with high resolution:

(in the order of 3 - 10 Hz.), direct estimation of bandwidth from the spectrum will usually fail due to the effect of the periodicity which causes zeros (or oscillations) in the spectrum and the effect of the glottal wave and radiation term (which combine to act as a low pass or smoothing filter). The work of Dunn* which is widely quoted, depended upon fitting to the spectrum templates that had impulse responses identical to the exponentially damped sinusoids used as sections of the formant model. Application of the wave-function analysis approach to filtered vowel sounds suggests a different type of mathematical "template fitting." The parameters which describe the envelope fit are the A, S, and C parameters. "A" describes the envelope peak amplitude; S, the envelope spread (approximate time interval containing the energy of the wave-function; and C, the location in time of the envelope peak. If a filtered region is obtained which has one and only one formant present, the effective damping of the time domain segment is strongly correlated to the actual bandwidth of the formant within that region.

A simple estimate of B_n can be obtained from the wave-function parameters by consideration of the following mathematics.

A single resonator can be described in the frequency domain as

$$Y(s) = \frac{2\pi F_n}{(S + \pi B_n)^2 + (2\pi F_n)^2} \quad (C-23)$$

With the time domain equivalent as

$$y(t) = \exp(-\pi B_n t) \sin(2\pi F_n t)$$

At the positive peaks corresponding to times t_p and t_{p+q}

$$y(t_p) = \exp(-\pi B_n t_p)$$

* Dunn, H.K., "Methods of Measuring Vowel Formant Bandwidths," JASA, Vol. 33, No. 12, December 1961, pp. 1737 - 1746.

At t_{p+q} , $y(t_{p+q}) = \exp(-\pi B_n t_{p+q})$. Therefore, B_n can be obtained by dividing the above equations and taking the logarithm of each side. This results in

$$B_n = \frac{1}{\pi(t_{p+q} - t_p)} \ln \left[\frac{y(t_p)}{y(t_{p+q})} \right] \quad (C-24)$$

This suggests that the amplitude parameters $A(1,n)$ and $A(i,n)$ of two wave-functions, and their corresponding time separation $C(i,n) - C(1,n)$ might be substituted into this last equation to produce bandwidth estimates of the form

$$\hat{B}_{n,i} = \frac{\ln [A(1,n)/A(i,n)]}{\pi[C(i,n) - C(1,n)]} \quad (C-25)$$

Several filtered segments (where formants are isolated) are shown in the figures with the corresponding envelope estimations generated by using the AS, and C parameters of the wave-function representation.

These waveforms suggest a bandwidth estimator composed of fitting exponential curves through the envelope peaks of the various wave-functions using a starting point of $A(1,n)$ located at $C(1,n)$ and then averaging in some form the estimated bandwidths. The simplest estimator in terms of the wave-function parameters in the sense that each estimate $\hat{B}_{n,i}$ is given equal weight is thus

$$\hat{B}_n = \frac{1}{\pi(M_n - 1)} \sum_{i=2}^{M_n} \hat{B}_{n,i} \quad (C-26)$$

Figures C-44 through C-46 show the impulse response of the corresponding vocal tract model section with the estimated F_n and B_n parameters vs. the overall wave-function representation.

Formant amplitudes are the least important of the formant parameters. Fant* has shown that by knowing the formant frequencies and bandwidths, the spectral amplitudes can be related along with the amplitude of the vocal tract impulse response. A physical verification that this is possible can be obtained by designing a cascaded vocal tract synthesizer such as the DVS. With this type of model, formant amplitude is not even specified. However, for the sake of completeness, an estimate of spectral amplitude in terms of wave-function parameters has been derived. If \hat{F}_n is the formant frequency estimate from the wave-function parameters, the formant amplitude estimate can be made from

$$\hat{A}_n = \left| \sum_{i=1}^{M_n} \psi [\Omega(i,n), jF_n] \right| \quad (C-27)$$

where ψ is defined as earlier.

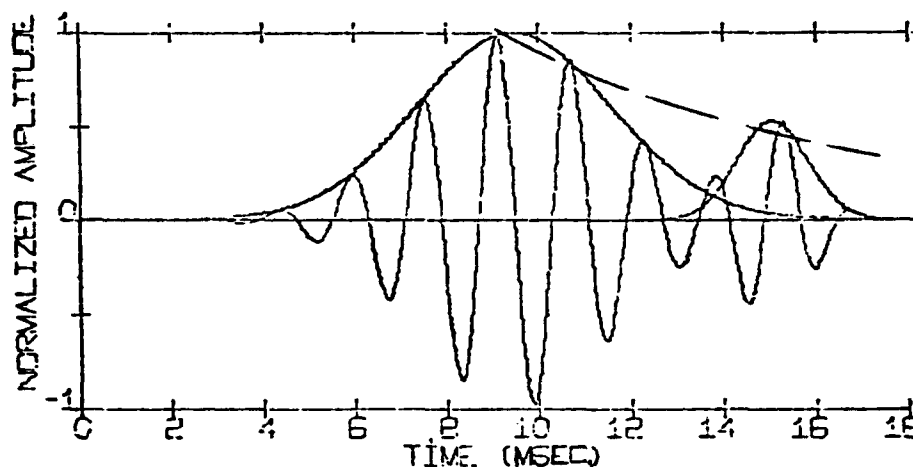


Figure C-41 Isolated pitchsegment for /A/ $F_1(0,.9)$ KHz. Dotted lines indicated effective damping.

* Fant, C. G., "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies," 1956, pp. 109, 120.

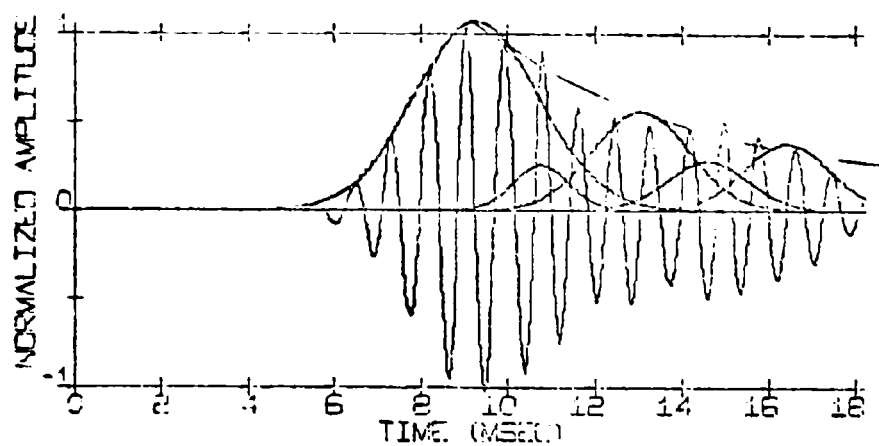


Figure C-42 Isolated pitch segment for Δ/R_2 (.9, 2) KHz.
Dotted lines indicate effective damping.

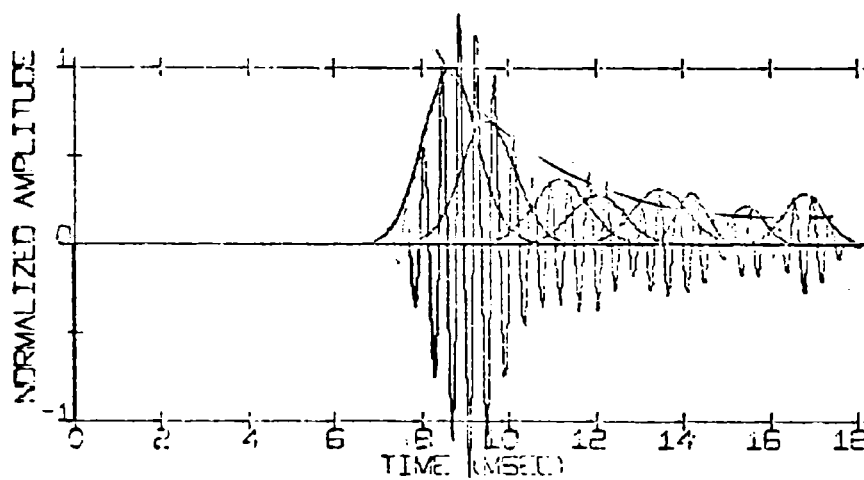


Figure C-43 Isolated pitch segment for Δ/R_3 (2, 3) KHz.
Dotted lines indicate effective damping.

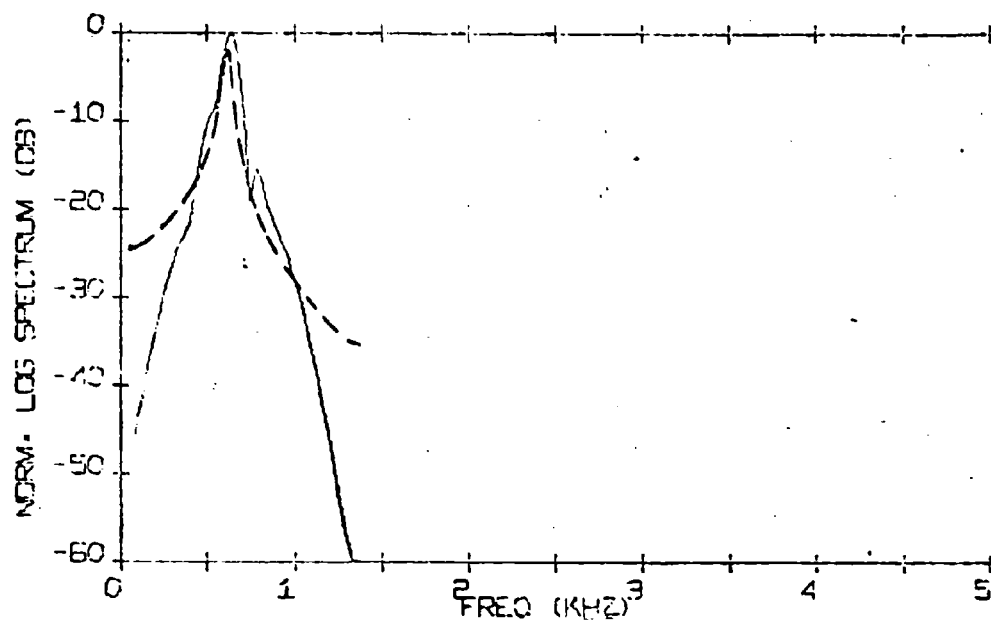


Figure C-44 Spectrum of wave-function representation (R_1 of $A/$).
(Solid versus resonator response using \hat{F}_1 , \hat{B}_1 (dotted).

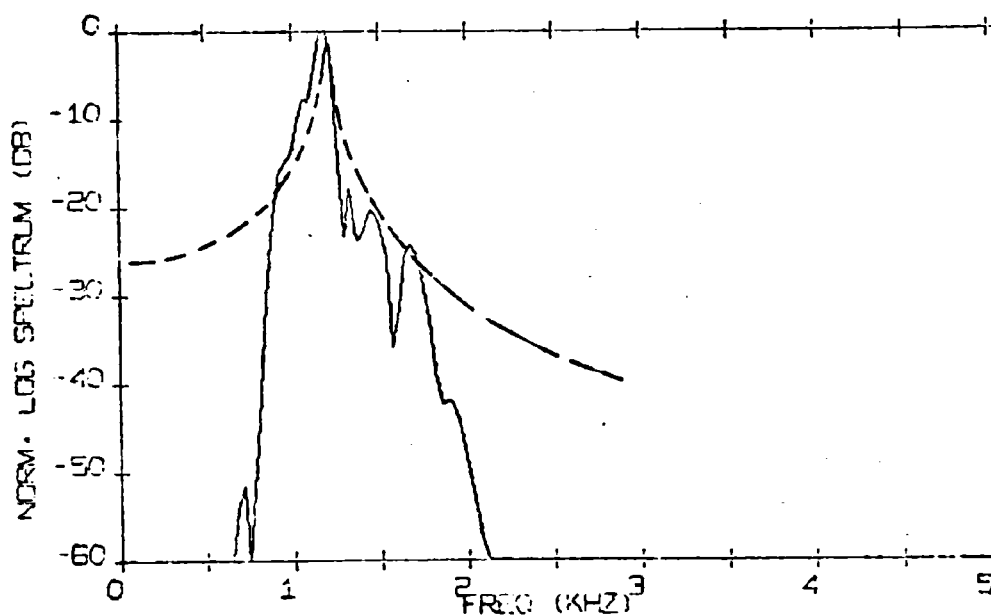


Figure C-45 Spectrum of wave-function representation (R_2 of $A/$)
(Solid) versus resonator response using \hat{F}_2 , \hat{B}_2 (dotted).

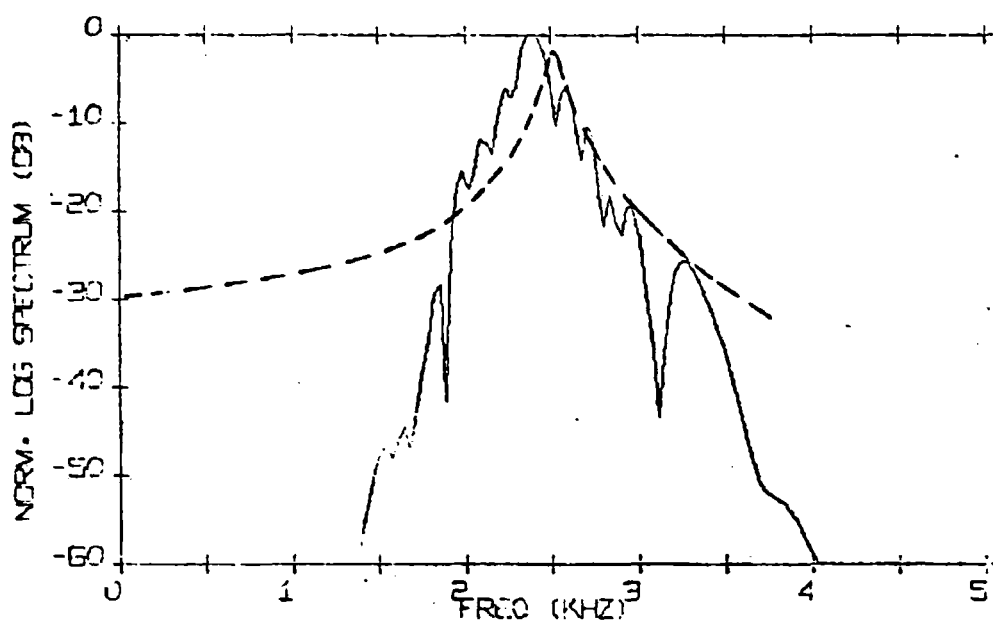


Figure C-46 Spectrum of wave-function representation (R_3 of Λ')
(Solid) versus resonator response using \hat{F}_3, \hat{B}_3 (dotted).

To check the validity of the proposed interrelationships, it was decided to generate ten different synthetic vowels and perform a complete analysis upon each to determine the estimate of the vocal tract model parameters. The vowels chosen were those of the Peterson-Barney study* with the average formant frequency values they suggested. The bandwidths were determined by the equation

$$B_n = (45 + 5F_n^3) 10^{-3} (\text{KHz.}) \quad 0 < f_n < 3 \text{ KHz.}$$

which is a good fit to the bandwidth data presented by Dunn over the region (0,3)KHz.

As previously discussed, the formant frequency is of most importance in the specification of vowel sounds. Over the ten vowels, the maximum error in the estimation of F_1 was 105 Hz. for the vowel 'ae/'. In the F_2 region the maximum error was 80 Hz. again for 'ae/'. For the F_3 region, /i/ had the maximum error of 410 Hz. These results are displayed for the ten vowels used in Figure C-47. The large errors in the F_3 region are believed due to the sampling rate of the system. Certainly as the effective frequency of the signal is increased, the measurement errors will increase. These errors can be reduced by either increasing the system sampling rate or by interpolating between points to estimate the extrema locations, or possibly by both methods. With these modifications, measurement accuracy of the F_3 region should be as good as the F_1 region. The use of an interpolative method to determine the F parameter was investigated, and the results show a marked reduction in error for region three. The maximum error in formant

* Peterson, G.E., and H. L. Barney, "Control Methods Used in a Study of the Vowels," JASA, Vol. 24, 1952, pp. 175-184.

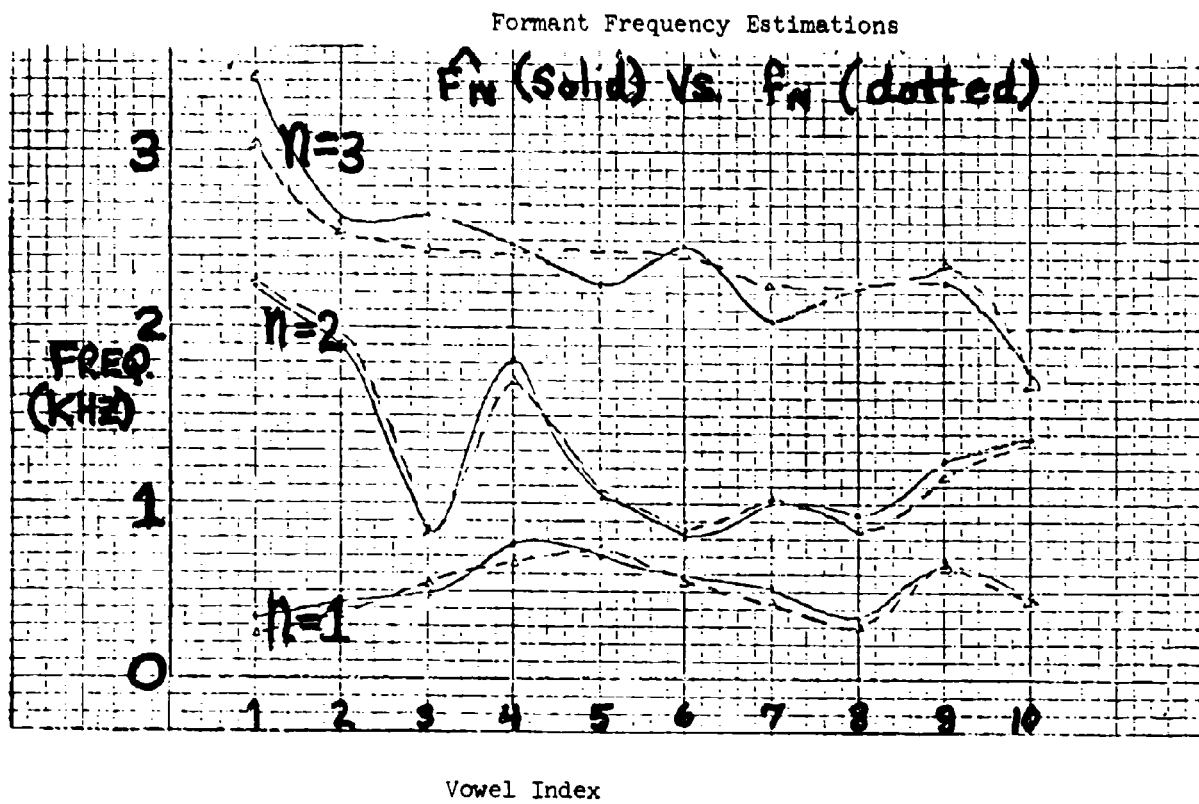


Figure C-47

frequency for the three bands were then 80 Hz., 65 Hz., and 65 Hz. respectively. This technique also gave a marked reduction in the standard deviation of the errors for the third region.

The maximum errors in the bandwidth estimation for the three regions were 30 Hz., 28 Hz., and 35 Hz. Although the percentage errors in the bandwidth estimations are quite large in some cases (84% error for Region 1 of /U/), the important factor is absolute error. Flanagan* has discussed difference limens (just perceivable differences) for formant frequencies, bandwidths, and amplitudes. He mentions that the difference limen for formant frequency appears to be 3 - 5% while that of formant bandwidth appears to be 30 - 40%. It is suggested that Flanagan's bandwidth limen is meaningful only in the sense that amplitude is also being allowed to vary. As stated by Flanagan, a 30 - 40% change in bandwidth causes roughly a 1.5 db amplitude change, which just happens to also be the difference limen for formant amplitude. It is postulated that, if both formant amplitude and frequency are held constant (as could be done with a parallel analog synthesizer), 20 - 30 Hz. deviations on B_1 (corresponding to percentage errors of up to 67%) would be totally imperceptible.

In the results presented so far, the assumption has been made that each region would consist of one and only one formant. Although this appears to be a realizable assumption for many vowels, it is rather doubtful that the assumption is valid for back vowels such as /a/ and /O/. From the Peterson-Barney data $F_2 - F_1$ equals 270 and 360 Hz. for /O/ and /a/ respectively. A

* Flanagan, J.L., "Speech Analysis Synthesis and Perception", Springer-Verlag, Berlin, Germany, 1965.

modification of the algorithm was developed which estimates both formant frequencies and assigns a mean bandwidth estimate to each of the formants. Results for /O/ and 'a' are shown in Table C-6.

VOWEL	F_1	\hat{F}_1	B_1	\hat{B}_1
/a/	730	696	47	55
/O/	570	527	46	45
VOWEL	F_2	\hat{F}_2	B_2	\hat{B}_2
/a/	1090	1056	51	55
/O/	840	817	48	45

Table C-6 Estimates of formant frequencies and bandwidths for closely spaced formants contained within a single region.

The results are extremely good with a maximum formant frequency error of 43 Hz. for \hat{F}_1 of /O/ and a maximum formant bandwidth error of 8 Hz. for \hat{B}_1 of /a/. Certainly these results should not be generalized to imply that accuracy of this order could be obtained for all situations. However, it is believed that the technique for extracting the parameters when two closely spaced formants are contained within a single region is generally valid.

Speech Project, Software and Hardware

The software and hardware development since the last technical report has been devoted to continued implementation of the SEL 810B Speech Analysis Laboratory and to continued work on the Video-to-Digital Converter (VIDIG) software for the Biological Sciences Department. There were three main aspects to this program which are as follows:

(1) Completion of Video-to-Digital Converter hardware system, and definition of the hardware/software complex to provide an on-line system for biological research.

(2) Continuation of SEL-810B software development including completion of all of Phase II except the mathematical operations for levels I and II which are currently under implementation.

(3) Completion of design of SEL-810B interface to speech station. The above items are discussed in more detail in the following paragraphs.

(a) On-Line System for Biological Research

The next phase of the project with the Video-to-Digital Converter is to generate the software to complete the combination of hardware and programming forming a unique tool for biological research.

In a joint meeting between the engineering and biological research people involved in this project, a set of goals defining the useful biological parameters to be extracted from the biological On-Line System were laid out. An outline of these goals is as follows:

1. Generalized input program for biological data from the video source using the VIDIG.
2. Establish scale to determine the ratio of internal (computer) units to external units.
3. Display descriptor words.
4. Sort and find bugpaths
5. Number of label bugpaths for display
6. Display video or path data
7. Connect bugpaths extrapolating across local voids of data

8. Eliminate unwanted bugpaths

9. Path Dynamics - for individual bugpaths and ensemble averages of all bugpaths

a) x velocity = $\frac{dx}{dt}$, y velocity = $\frac{dy}{dt}$

b) linear velocity $\bar{V}(t) = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} e^{i\theta(t)}$

$\theta(t)$ = direction of travel

c) curvature, radius of curvature

d) length of paths = $\theta \sqrt{\frac{dx^2}{dt} + \frac{dy^2}{dt}}$

e) turning = $\int_0^P \frac{d\theta}{dt} dt = \theta(P) - \theta_0$

f) path haiti = length of time in position

The physical system as it exists on the IBM 1800 computer is shown in Figure C-48 in the form of a block diagram.

Note that the Video-Digital Converter is capable of quantizing and rendering to a computer any televisable event. The light pen input is accomplished by focusing a television camera onto a clear plastic screen and a small pen-light flashlight serves as the televisable event. These coordinates are read into the computer then averaged and can be displayed in non-store mode on the display scope as a pointer to data structures.

The on-line system for biological research is being developed as a stand-alone single station on-line system for the 1800 computer similar in structure to that existing on the 360 computer. The on-line system operates with the disk, occupies about $\frac{1}{4}$ of the 16K of core leaving the rest for data and has full alpha-numeric and curvilinear display available on the Tektronics 611 display scope.

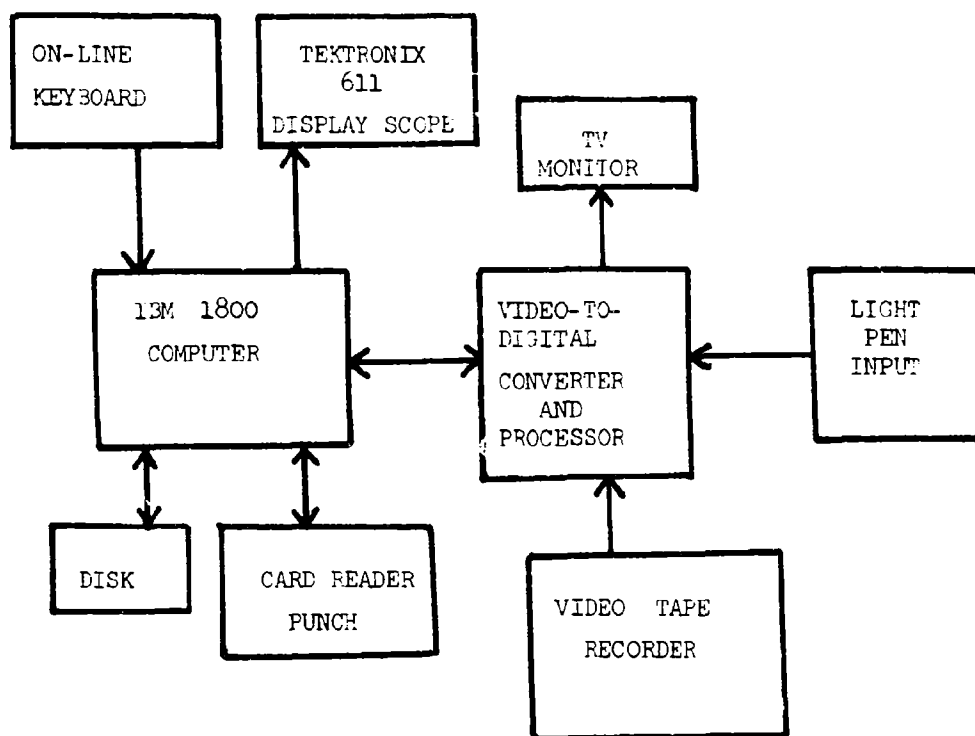


Figure C-48 VIDIG On-Line System for Biological Research

The Biological System will define two levels of this on-line system: Level VI for the VIDIO input, display and special programs necessary for defining and operating on the three dimensional data structures unique to the VIDIO. Level V will be for the display and further analysis of the biological data after processing and refinement of Level VI. Data is manipulated with operator control via the keyboard input. Each operation calls the macro controller program which allows for two alphabetic and three numeric trailing predicates and is terminated by pushing the "Return" button. For example, the "Display" operator on Level VI for viewing the quantized video frames stored in core has the following format:

DISPLAY $\alpha_1 \alpha_2 N_1, N_2, N_3$ RETURN

where

α_1 = D display in Dot mode

L display Line connecting the dots

α_2 = N print the page number at the starting coordinate of each page

N_1 = starting page number to be displayed

N_2 = number of pages to be displayed

N_3 = delay count for slowing down the display if desired

The operation is defined if none of the trailing predicates are specified (i.e. Display Return). This will display all of the video pages in dot mode with no delay.

Currently the Level VI programs most of which are disk based are well underway and may be broken into three broad categories:

1. Input/Output Programs

- a. VDP Data Input Program: This program initialized the data input

channel to read in the VDP at maximum rate and defines the data structure to be used.

b. Load/Store is for transferring data within core or to and from portions of the disk allotted for data storage.

c. Display for viewing quantized video data in core or for displaying the "bugpaths" on data that has been processed.

d. Initialize light pen input: This program initializes the digital input channel to read 16 coordinates from the VIDIG, average and display in non-store mode the non-zero coordinates in the list and makes available these coordinates to other programs. This program then continues to run when the computer is idle until the Reset button is pushed.

2. Special Purpose Programs

a. Find Bugpaths: This program sorts through the video data which is read in on a frame by frame basis searching for possible bugpaths that link together in time and space for later calculation of path dynamics.

b. Ensemble averaging of path dynamics is necessary for statistical averages of such parameters as velocity, changes in direction, net displacement, and rotation before and after the application of a stimulus.

c. Evaluate descriptor words: At the beginning of each new frame of video data, the VIDIG sends to the computer a descriptor word which contains an encoding of the current scan rate and four bits of information conveying the on-off status of four possible stimuli applied to the biological organisms under study. This program displays this information on the display scope in a mode specified by keyboard control.

d. Find Centroids: This program calculates the centroids of points

that lie within a user specified mask about each coordinate in the data. This can then be used to replace the outlines of organisms with their centroid point for path dynamics calculations.

3. Mathematical Operations

These operations will be fixed point routines to calculate those mathematical parameters useful in biological research within the defined data structures.

Double predicate operations: $\oplus, \ominus, \odot, \oslash$

Single predicate operators: Square, square root, central difference, central sum, arctangent.

(b) SEL-810B Software

The software development program for the SEL-810B which was outlined in the Thirteenth Quarterly Report is about two months behind the proposed target date of June 31, 1970. As of this writing Phase II of the software development is nearing completion and it is anticipated that Phase II should be operational early in August.

Following the completion of Phase II the speech system software (Phase III) will be added to the SEL-810B system. Since the speech analysis/synthesis software is now well defined a fully operational speech analysis/synthesis system should be available on the SEL-810B by the end of the summer. The present status of the SEL software system is summarized in the following paragraphs.

The Machine Language Disk Controller

A routine was written in SEL 810 machine language for the purpose of loading and storing programs such as the assembler on disk. This routine accepted a block number in the switch register and loaded or stored all

core starting at that block; it was of utility in generating the SEL 8103 operating system.

The SEL 8103 Operating System

An SEL operating system was designed to provide a set of basic operators for loading and storing programs on disk, linking various programs together, and executing programs resident in core. This routine accepts operator commands from the SEL teletype keyboard.

The SEL Mnembler

The SEL Mnembler two pass assembler was modified to make use of the operating system routines; the assembler reads cards from the card reader, makes the two passes from disk, and writes relocateable object output on disk.

The SEL Relocateable Loader

The SEL Relocateable Loader was modified to read relocateable object output produced by the assembler from disk and load it into core.

Construction of the SEL 810B On-Line Speech Acquisition and Analysis System

During this reporting period the basic controllers for an SEL 8103 two-station on-line system have been developed.

Input Keyboard Interrupt Processing Routine and OPCON

The input keyboard interrupt processing routine and operational controller (OPCON) have been fully written and checked out. These include the routines for processing of console programs (USER) and repeating buttons (REPEAT). Average button processing time (overhead) has been measured to be 90 μ sec.

Display Generation

Display generation routines have been implemented to provide character

display and line drawing. The character generation routines provide a variable character size. The typing level operations have been implemented. The curvilinear display program accepts two lists of integers normalized with a scale of 1024.

Debugging Level

The debugging level has been implemented on on-line debugging ease. This provides a set of software registers, operations to evaluate and modify core, read and write disk, and an on-line assembler.

Data Structure Manipulation Routines

A universal data structure has been developed for loading and storing variable length blocks of data. This structure is used to load disk-based routines into core, perform necessary relocation, update and load user programs, floating point lists (Level II data), speech data, etc. Routines provided to manipulate the data structure are load item, update item, and repack item (which is the structure's "garbage collector").

Storage Allocation

A memory paging scheme has been devised for main storage management. Elements in the data structure are assigned priorities in their competition for main storage. When main storage is completely used, the element with minimum priority is purged from core in the attempt to allocate space.

The Console Program Generation

The console program generation (LIST) routine has been implemented for console program building, editing, and storage.

Assembler for 360/75

An assembler for the SEL 810B has been written for the 360/75. Its

purpose is to provide an assembly language listing on the 360 line printer with optional object output on punched cards.

The Numerical Operators

The numerical operators on the integer level have been implemented. The floating point arithmetic routines add, subtract, multiply, and negate have been written and checked out. The floating point data format for the SEL 810B has been chosen to contain 32 bits of mantissa and 16 bits of characteristic (scale of 2). The mathematical levels I and II (floating point single number operations, vector operations respectively) are currently being implemented. Additional levels for speech synthesis/analysis, and phoneme manipulation are being designed.

(c) SEL-810B Interface to Speech Station

An on-line keyboard and Tektronix 611 kisplay scope have been incorporated into the SEL hardware speech system and are presently being used for the completion of Phase II of the software system. Since D/A converters are needed to drive the 611 display scope, these have also been added to the system. The present hardware configuration is depicted in block diagram form in Figure C-49.

The present hardware configuration is characterized by the restriction that all data transfers to and from the SEL must be made under direct program control rather than under the supervision of either of the two available Block Transfer Control Units contained within the SEL. Three devices are daisy chained to the SEL in the present system. These are, a teletype, a card reader, and a four channel I/O multiplexor two of which are used. One channel provides a data link between the SEL and the RW-400 which has been

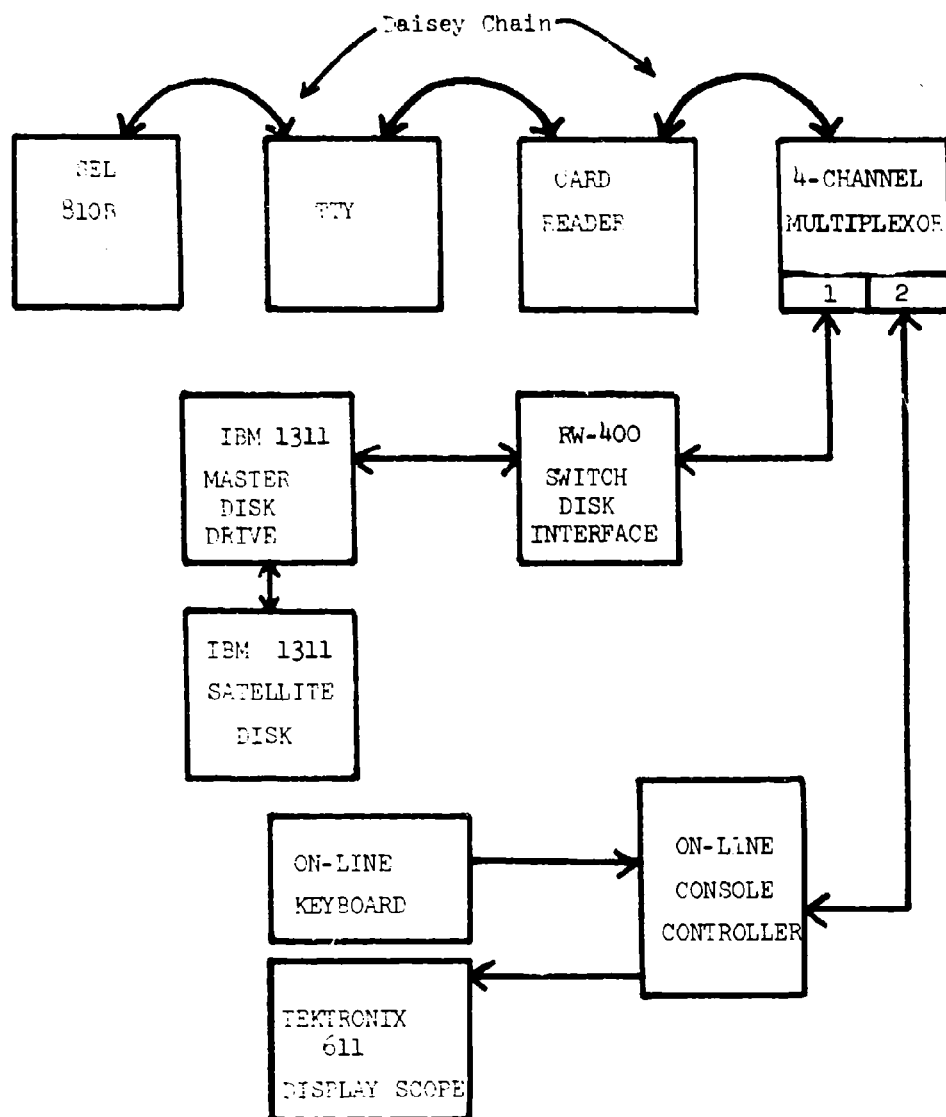


Figure 2-49 Present SEL-810B hardware configuration.

modified to serve as a disk controller for the IBM 1311 disk drive and satellite. The other channel of the multiplexor is tied to an on-line console controller. The input side of this controller sends a process interrupt to the SEL every time a keyboard button is pressed. The SEL responds to the interrupt by executing a digital input from the controller and thus reading the button code.

The output side of the on-line console controller drives a Tektronix 611 storage scope. 10 bits of each output are DAC data. One bit specifies whether the x or y DAC is to be loaded. One bit triggers the unblank one shot and another bit is for erase.

The SEL-810B hardware system which is presently under development is shown in Figure C-50. The design of the system has been completed and is currently under construction. The system is characterized by the capability for two simultaneous block transfer I/O operations. In addition the FW 400 Switch/disk controller, will be replaced by a disk controller whose design has been optimized for the problem of recording long intervals of speech. This device will facilitate the use of the SEL core as a double buffer. Sydirchs*, a SYstem for DIgitally REcording Human Speech, will execute input block transfers while the disk controller is executing output to the 1311 disk memory.

There are two additional I/O control devices essential to the system. One is a two station On-Line Interface and the second a Synchronous Nanohumper Sampler. This latter device is programmed from the on-line console to sample the output of a Nanohumper at fixed intervals of time as prescribed by the on-line user. This allows the Nanchumper to serve as a conventional

* Described in the 13th Quarterly Report

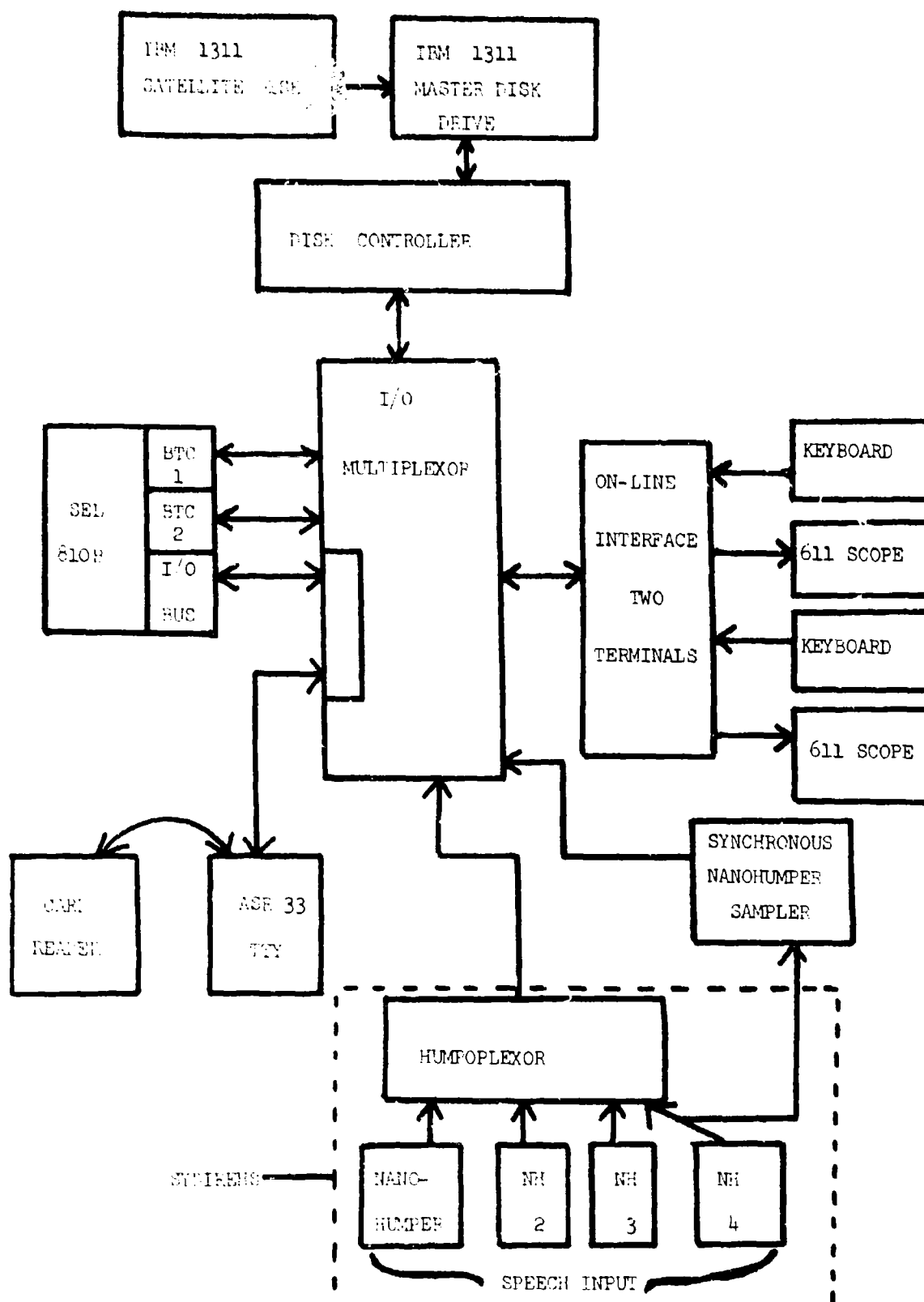


Figure 2-10 Proposed SEL-810B Hardware System

A/D converter with selectable sample rates. Thus it is seen that the system has four essential hardware I/O device controllers: the Disk Controller, SYDIPHS, the On-Line Interface, and the Synchronous Nanohumper Sampler. In order to accomodate the I/O devices and to leave room for future expansion of the hardware system, a versatile I/O multiplexor has been proposed and is presently under construction.

To the device side of the multiplexor sixteen device controllers are connected. To the computer side three distinct data transfer facilities are connected. Two of these facilities are the Block Transfer Control units or BTC's. The third facility is the standard I/O control set and the data bus. All data is actually transferred over the data bus. However, from the viewpoint of an external device, it is as though that device can be tied to any one of three data channels. Any device may transfer data via any one of the three data transfer facilities. A device is "tied" to one of the BTC's by means of a command issued over the standard I/O control set.

Once this tie is made, the device will remain tied to that BTC until the entire data block has been transferred. Responding to the device's data transfer request, the BTC will grant each request according to priorities among the other BTC and the standard I/O control set. The I/O Multiplexor enables the SEL to transfer data to or from any two devices in block mode while communicating with any of the fourteen remaining devices in the Direct Program Control mode using the Standard I/O control set. (Direct Program Control means that a full I/O instruction is executed for each data word transferred.)

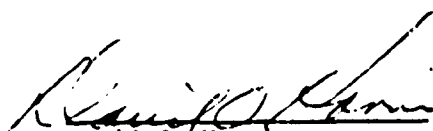
The Disk Controller will transfer data via Block Transfer Control.

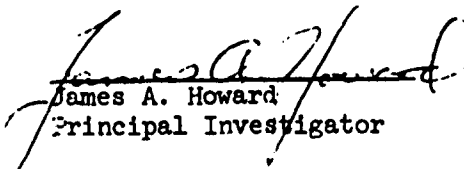
Consequently, only one I/O instruction is required to transfer an entire block of data. When a block transfer begins, the SEL provides the controller with the starting sector address and disk surface. These parameters are automatically incremented if the size of the data block requires. At the end of the transfer, a disk status word is presented to the SEL. It identifies the final sector address and the final disk surface. This feature facilitates the recording of a full cylinder of speech data with minimal program intervention.

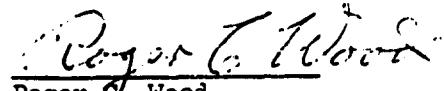
Conclusion

Tasks established in the original contract have been accomplished. Research to date has produced reliable end-products which are being successfully employed in computer technology; it has also produced other promising factors which warrant further exploration particularly in the areas relating to computer networks and the communication between humans and computers.

Submitted By:


David O. Harris
Principal Investigator


James A. Howard
Principal Investigator


Roger O. Wood
Principal Investigator

Unclassified

DOCUMENT CONTROL DATA - R & D		
(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)		
1. ORIGINATING ACTIVITY (Corporate author)		20. REPORT SECURITY CLASSIFICATION
University of California Santa Barbara, California 93106		Unclassified
		25. GROUP
3. REPORT TITLE		
RESEARCH IN ON-LINE COMPUTATION		
4. DESCRIPTIVE NOTES (Type of report and inclusive dates)		
Scientific. Final. 26 April 1966 - 30 June 1970		
5. AUTHOR(S) (First name, middle initial, last name)		
David O. Harris James A. Howard Roger C. Wood		
6. REPORT DATE	70. TOTAL NO. OF PAGES	70. NO. OF REFS
30 June 1970	117	7
80. CONTRACT OR GRANT NO. ARPA Order No. 865 AF19(628)-6004		80. ORIGINATOR'S REPORT NUMBER(S)
8. PROJECT NO. 8684 n/a n/a		
9. DOD ELEMENT 61101D		80. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
9. DOD SUBELEMENT n/a		AFCRL-70-0535
10. DISTRIBUTION STATEMENT		
1 - This document has been approved for public release and sale; its distribution is unlimited.		
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY
This research was supported by the Advanced Research Projects Agency.		Air Force Cambridge Research Lab. (LR) L. G. Hanscom Field Bedford, Massachusetts 01730
13. ABSTRACT		
<p>The report covers the on-line computing system development from 1966 through 1970. It includes a general resume of progress through December, 1969 and a detailed progress from then through June 30, 1970. The improved version of the on-line system substantially improves system reliability and presents users new options. Significant progress in speech analysis/synthesis project includes: improved techniques for deriving accurate data from ASCON parameters, good results from the steady-state vowel recognizer, and one-pass analysis and synthesis. The 1800 has been improved so that it is a more effective research tool supporting the speech effort.</p>		

DD FORM 1473
NOV 65

Unclassified

Security Classification

Unclassified

Security Classification

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Network Control Program						
Improved System Software						
Multi-Teletype Controller						
High Speed Data Buffer						
Gaussian Cosine Modulation Model						
Adaptive Filtering						
Fixed Filtering						
Sinusoids						
Preprocessing Acoustic Waveforms						
Wave-Function Analysis						
Extrema Detection/Correction						
Mapping Formant Frequencies						
Steady-State Vowel Recognition						
Recognition of Phonemes						
Data Compression Studies						
SEL 810B Software						
SEL 810B Speech Interface						
OLS Biological Research						

Unclassified

Security Classification